Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Flexible and Fast Spatial Return Level Estimation Via a Spatially Fused Penalty

Danielle Sass, Bo Li & Brian J. Reich

To cite this article: Danielle Sass, Bo Li & Brian J. Reich (2021): Flexible and Fast Spatial Return Level Estimation Via a Spatially Fused Penalty, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2021.1938584

To link to this article: https://doi.org/10.1080/10618600.2021.1938584

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2021, VOL. 00, NO. 0, 1–19
https://doi.org/10.1080/10618600.2021.1938584

Taylor & Francis
Taylor & Francis Group

Flexible and Fast Spatial Return Level Estimation Via a Spatially Fused Penalty

Danielle Sass[a], Bo Li[a], and Brian J. Reich[b]

[a]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL; [b]Department of Statistics, North Carolina State University, Raleigh, NC

ABSTRACT

Spatial extremes are common for climate data as the observations are usually referenced by geographic locations and dependent when they are nearby. An important goal of extremes modeling is to estimate the T-year return level. Among the methods suitable for modeling spatial extremes, perhaps the simplest and fastest approach is the spatial generalized extreme value (GEV) distribution and the spatial generalized Pareto distribution (GPD) that assume marginal independence and only account for dependence through the parameters. Despite the simplicity, simulations have shown that return level estimation using the spatial GEV and spatial GPD still provides satisfactory results compared to max-stable processes, which are asymptotically justified models capable of representing spatial dependence among extremes. However, the linear functions used to model the spatially varying coefficients are restrictive and may be violated. We propose a flexible and fast approach based on the spatial GEV and spatial GPD by introducing fused lasso and fused ridge penalty for parameter regularization. This enables improved return level estimation for large spatial extremes compared to the existing methods. Supplemental files for this article are available online.

ARTICLE HISTORY
Received November 2019
Revised May 2021

KEYWORDS
Fused lasso; Fused ridge; Generalized extreme value distribution; Generalized Pareto distribution; Spatial extremes

1. Introduction

Extreme value analysis is important because it allows us to assess and quantify the extent to which a rare event could occur. This is especially relevant in environmental sciences where extreme natural events such as hurricanes, wild fires, tornadoes, and floods can cause lasting human impacts. Classifying extreme values is commonly done in two ways: block maxima and peaks over threshold. The block maxima approach divides the dataset into equal periods and chooses the maximum value from each period. The generalized extreme value (GEV) distribution is the limiting distribution of block maxima and is known to fit accurately for large blocks (Fisher and Tippett 1928; Gumbel 1958; Coles 2001). The block maxima approach can be wasteful as it only chooses one point from each period. In the peaks over threshold approach, a threshold limit is chosen and all points exceeding the limit are selected to form the extreme dataset. For a high enough threshold, the generalized Pareto distribution (GPD) arises as the limiting distribution to model data over the threshold (Pickands 1975; Davison and Smith 1990).

An important goal of extremes modeling is to estimate the T-year return level, which is highly related to the marginal behavior of extremes. The T-year return level is the value expected to be exceeded on average once every T years. The estimation of return level is essential in the study of environmental science. For example, the return level estimation of sea levels is useful for the management of coastal zones.

Extremes in environmental data are typically observed over a spatial domain and thus the spatial variation and dependency of extremes are also exhibited in such data. Many techniques for modeling environmental extreme data have been developed that focus on either the site-wise marginal behavior or the spatial dependence of the extreme values. Characterizing the marginal distribution is often done through latent variable models (Coles and Casson 1998; Casson and Coles 1999; Cooley, Nychka, and Naveau 2007; Sang and Gelfand 2009). Latent variable models introduce spatial variation through Gaussian processes of the parameters of the marginal distribution. This is often accomplished via a hierarchical specification of the joint distribution followed by Bayesian inference which is computationally intensive and dependent on prior distributions. Besides latent variable models, the spatial GEV and spatial GPD are perhaps the simplest and fastest approach to modeling the spatial variation of marginal distributions. These methods assume marginal independence and only account for dependence by allowing the parameters to vary spatially through linear functions.

However, if the primary interest is in identifying how the values at each site are related to each other, the modeling of spatial dependence will become the focus. Max-stable models (Smith 1990; Davis and Resnik 1993; Schlather 2002) have been popular choices for capturing the dependence structure. Recently, more advances have emerged beyond max-stable models (Wadsworth and Tawn 2012; Huser, Opitz, and Thibaud 2017, 2021; Huser and Wadsworth 2019; Castro Camilo and Huser 2020). For instance, a new development proposes a flexible Bayesian model that allows for varying spatial dependence (Bopp, Shaby, and Huser 2021). For a more comprehensive review of recent

CONTACT Danielle Sass dsass2@illinois.edu Department of Statistics, University of Illinois at Urbana-Champaign, 725 South Wright Street, Champaign, IL.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

research on modeling spatial extremes, see Davison et al. (2012) and Davison et al. (2019).

In this article, we develop computationally efficient yet flexible fused spatial GEV and spatial GPD models through penalized regression. Our fused spatial extremes models share the advantage of the traditional spatial GEV and GPD models for being applicable to large data sets, however, our models allow for a nonparametric representation when describing the spatial variability of the GEV and GPD parameters and thus largely improves the return level estimation. Compared to max-stable models that demand a dependency structure, our fused spatial extremes models do not require such specifications but rather regulates spatial variability through the penalties. Compared to Bayesian implementation of max-stable models that are usually computationally intensive for handling large covariance matrices, our methods dramatically improve the computational efficiency. Furthermore, the proposed models require no stationarity assumption which is commonly made in modeling spatial extremes.

2. Review of Spatial Extremes Models

We begin by introducing the univariate GEV distribution. Let M_n represent the maximum of a sequence of independent random variables, $X_1, \ldots, X_n$, which have common distribution F. If there exists sequences of constants $\{a_n\} > 0$ and $\{b_n\} \in \mathbb{R}$ such that $\frac{M_n - b_n}{a_n}$ has a nondegenerate limiting distribution G as $n \to \infty$, then G must be a member of the GEV class with distribution function

$$G(y) = \begin{cases} \exp[-\{1 + \frac{\xi}{\sigma}(y-\mu)\}_+^{-1/\xi}], & \text{if } \xi \neq 0 \\ \exp[-\exp\{-(y-\mu)/\sigma\}], & \text{if } \xi = 0, \end{cases} \quad (1)$$

where $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$ represent the location, scale, and shape parameters, respectively (Coles 2001). The GEV distribution is max-stable, which means the maximum of n independent and identically GEV(μ, σ, ξ) distributed variables follows a GEV distribution with the same shape parameter ξ but different location and scale parameters, that is,

$$G^n(a_n y + b_n) = G(y), \quad y \in \mathbb{R}, n \in \mathbb{N}.$$

An alternative to modeling the block maxima is to model the excesses above a high threshold using the GPD distribution. For a large enough threshold u, the conditional distribution function of the excesses $Y = X - u | X > u$ is approximately

$$\begin{aligned} H(y) &= P(X - u > y | X > u) \\ &= \begin{cases} 1 - (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - \exp(-\frac{y}{\tilde{\sigma}}), & \text{if } \xi = 0. \end{cases} \end{aligned} \quad (2)$$

The limiting model for threshold exceedances, Equation (2), corresponds to fitting (1) to maxima with $\tilde{\sigma} = \sigma + \xi(u - \mu)$ and an identical ξ. Since our models are extensions of the spatial GEV and spatial GPD, we will review those two methods in more detail. Review of an example of a max-stable model, Schlather (2002), a Bayesian model called hierarchical kernel extreme value process (HKEVP) (Reich and Shaby 2012), and a latent variable model (Davison, Padoan, and Ribatet 2012) are in Appendix A.

2.1. Spatial GEV Model

Let $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$ be the observed block maxima at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ over a spatial domain $\mathcal{D}$. Assume the marginal distribution of $y(\mathbf{s})$ is GEV with parameters $\mu(\mathbf{s})$, $\sigma(\mathbf{s})$ and $\xi(\mathbf{s})$. When standard maximum likelihood estimates are not available due to an intractability of the full likelihood, a composite likelihood is a natural alternative (Varin and Vidoni 2005; Sang and Genton 2014). The spatial GEV model assumes marginal independence given the GEV parameters and thus enables a composite log-likelihood approximation to the full log-likelihood (Lindsay 1988) given by

$$l\{\boldsymbol{\theta}(\mathbf{s}))\} = \sum_{i=1}^{n} l\{\boldsymbol{\theta}(\mathbf{s}_i); y(\mathbf{s}_i)\}, \quad (3)$$

where n denotes the number of sites, and $l\{\boldsymbol{\theta}(\mathbf{s}_i); y(\mathbf{s}_i)\}$ is the univariate GEV log-likelihood function at site $\mathbf{s}_i$ with parameter vector $\boldsymbol{\theta}(\mathbf{s}_i)$.

When $\xi(\mathbf{s}_i) \neq 0$, we have $l\{\boldsymbol{\theta}(\mathbf{s}_i); y(\mathbf{s}_i)\}$ in terms of $\mu(\mathbf{s}_i), \sigma(\mathbf{s}_i)$, and $\xi(\mathbf{s}_i)$ as

$$\begin{aligned} &l\{\mu(\mathbf{s}_i), \sigma(\mathbf{s}_i), \xi(\mathbf{s}_i); y(\mathbf{s}_i)\} \\ &\quad = -m \log\{\sigma(\mathbf{s}_i)\} - \left\{1 + \frac{1}{\xi(\mathbf{s}_i)}\right\} \\ &\qquad \times \sum_{t=1}^{m} \log\left[1 + \xi(\mathbf{s}_i)\left\{\frac{y_t(\mathbf{s}_i) - \mu(\mathbf{s}_i)}{\sigma(\mathbf{s}_i)}\right\}\right] \\ &\qquad - \sum_{t=1}^{m}\left[1 + \xi(\mathbf{s}_i)\left\{\frac{y_t(\mathbf{s}_i) - \mu(\mathbf{s}_i)}{\sigma(\mathbf{s}_i)}\right\}\right]^{-\frac{1}{\xi(\mathbf{s}_i)}}, \end{aligned} \quad (4)$$

provided that $1 + \xi(\mathbf{s}_i)\left\{\frac{y_t(\mathbf{s}_i) - \mu(\mathbf{s}_i)}{\sigma(\mathbf{s}_i)}\right\} > 0$ for $t = 1, \ldots, m$, where m is the number of replicates for each site. At parameter combinations for which at least one of the observed data falls beyond the bounds of the distribution, the likelihood is zero and the log-likelihood equals $-\infty$. The case $\xi(\mathbf{s}_i) = 0$ requires separate treatment using the Gumbel limit of the GEV distribution and is defined on $-\infty < y_t(\mathbf{s}_i) < \infty$ for $t = 1, \ldots, m$.

The composite log-likelihood models the site-wise marginal behavior independently and the spatial variability is usually imposed through the GEV parameters. The GEV parameters are commonly described through parsimonious regression models, which may be functions of space, environmental and other covariates, and random effects (Padoan, Ribatet, and Sisson 2010). For example, μ and σ may be assumed to be linearly related to coordinates of locations while ξ is assumed to be a constant (Coles and Tawn 1996), modeled as

$$\begin{aligned} &\mu(\mathbf{s}) = \boldsymbol{\beta}_\mu^T X_\mu(\mathbf{s}), \quad \log\{\sigma(\mathbf{s})\} = \boldsymbol{\beta}_\sigma^T X_\sigma(\mathbf{s}), \\ &\xi(\mathbf{s}) = \beta_{\xi_0}, \end{aligned} \quad (5)$$

where $X_\mu(\mathbf{s})$ and $X_\sigma(\mathbf{s})$ are vectors of covariates with coefficients $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\sigma$, respectively. Then maximum likelihood is used to fit the model to the observed data by plugging (5) into Equation (4) and optimizing with respect to $\boldsymbol{\theta} = \{\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \beta_{\xi_0}\}$. The reason ξ is assumed to be a constant is due to the challenge of its estimation and the sensitivity of the return level to this parameter.

Estimates of extreme quantiles of block maxima distributions are given in terms of the parameters. By inverting (1) and letting $G(y) = 1 - \frac{1}{T}$, we have the T-year return level r_T as

$$r_T(\mathbf{s}_i) = \begin{cases} \mu(\mathbf{s}_i) - \frac{\sigma(\mathbf{s}_i)}{\xi(\mathbf{s}_i)}\left[1 - \left\{-\log(1-\frac{1}{T})\right\}^{-\xi(\mathbf{s}_i)}\right], & \text{if } \xi(\mathbf{s}_i) \neq 0 \\ \mu(\mathbf{s}_i) - \sigma(\mathbf{s}_i)\log\{-\log(1-\frac{1}{T})\}, & \text{if } \xi(\mathbf{s}_i) = 0, \end{cases}$$

given one fits a GEV to an annual maximum of a stationary process. One needs to adjust T for different block sizes (Cao and Li 2018).

2.2. Spatial GPD Model

One disadvantage of modeling block maxima in extreme value analysis is that many extreme data are wasted simply because they are not the block maxima. As a result, threshold models have become a popular modeling alternative.

Having determined a threshold, the parameters of the GPD can be estimated by maximum likelihood. Let spatial data $y_1(\mathbf{s}_i), \ldots, y_{k(\mathbf{s}_i)}(\mathbf{s}_i)$ be the $k(\mathbf{s}_i)$ excesses of a threshold u at site $\mathbf{s}_i$ for $i = 1, ..., n$. The composite log-likelihood is again given by Equation (3), where n denotes the number of sites and $l\{\boldsymbol{\theta}(\mathbf{s}_i); y(\mathbf{s}_i)\}$ is the univariate GPD log-likelihood function at site $\mathbf{s}_i$ with parameter vector $\boldsymbol{\theta}(\mathbf{s}_i)$. When $\xi(\mathbf{s}_i) \neq 0$, we have $l\{\boldsymbol{\theta}(\mathbf{s}_i); y(\mathbf{s}_i)\}$ in terms of $\tilde{\sigma}(\mathbf{s}_i)$ and $\xi(\mathbf{s}_i)$ as

$$l\{\tilde{\sigma}(\mathbf{s}_i), \xi(\mathbf{s}_i); y(\mathbf{s}_i)\} = -k(\mathbf{s}_i)\log\{\tilde{\sigma}(\mathbf{s}_i)\} - \left\{1 + \frac{1}{\xi(\mathbf{s}_i)}\right\} \times \sum_{t=1}^{k(\mathbf{s}_i)} \log\left\{1 + \xi(\mathbf{s}_i)\frac{y_t(\mathbf{s}_i)}{\tilde{\sigma}(\mathbf{s}_i)}\right\}, \quad (6)$$

provided $\{1 + \xi(\mathbf{s}_i)\frac{y_t(\mathbf{s}_i)}{\tilde{\sigma}(\mathbf{s}_i)}\} > 0$ for $t = 1, \ldots, k(\mathbf{s}_i)$, otherwise $l\{\tilde{\sigma}(\mathbf{s}_i), \xi(\mathbf{s}_i); y(\mathbf{s}_i)\} = -\infty$. The case $\xi(\mathbf{s}_i) = 0$ requires separate treatment using the exponential limit of the GPD distribution and is defined on $0 < y_t(\mathbf{s}_i) < \infty$ for $t = 1, \ldots, k(\mathbf{s}_i)$.

Similar to the spatial GEV model described in Section 2.1, the spatial variability of the threshold excesses can be modeled through the GPD parameters. A popular choice is to assume $\tilde{\sigma}$ is log-linearly related to spatial covariates and ξ is a constant as defined in Equation (5). The m-observation return level r_m, provided m is sufficiently large to ensure that $r_m > u$, can be obtained by letting Equation (2) equal to $1/m$. The T-year return level is the level expected to be exceeded once every T years. If there are k observations per year, this corresponds to the m-observation return level with $m = Tk$.

3. Fused Spatial GEV and GPD Models

Due to their computational efficiency, the spatial GEV and spatial GPD can be applied to large spatial extremes data. Surprisingly, the return level estimation based on spatial GEV is shown to be comparable to various max-stable models (Cao and Li 2018). However, if the spatial domain of observations is vast, the constant shape parameter usually assumed in the spatial GEV and GPD will likely be violated and this may cause deteriorated return level estimation as return levels are very sensitive to even a small perturbation of the shape parameter. Furthermore, the location and scale parameters in those two models may not simply follow a parametric relationship with the available covariates.

In such cases, it would be more appropriate to account for the spatial heterogeneity of the shape parameter, as well as to allow for flexible forms of spatial variability in location and scale parameters. Unlike Bayesian methods that assume a latent Gaussian process for the parameters, we propose to regulate the shape parameter using the fused lasso or fused ridge penalty. To attain flexibility of other parameters that are usually restricted to a parametric, often linear, form in the spatial GEV and spatial GPD, we further propose to regulate all three parameters in the spatial GEV or two parameters in the spatial GPD using a fused lasso or fused ridge penalty. The fused penalty has been used to model the clustering pattern or smoothness of spatial data by penalizing differences in parameter estimations of nearby locations (Tibshirani and Taylor 2011; Parker, Reich, and Eidsvik 2016; Tansey et al. 2018; Li and Sang 2019). We will proceed by outlining the methodology only for the fused spatial GEV model, however, the fused spatial GPD can be analogously developed by replacing the likelihood and parameters in the spatial GEV with those from the spatial GPD.

3.1. Fused Penalties

To define the fused penalty in the context of spatial extremes, consider an undirected graph G consisting of n vertices and edge set $\mathcal{E}$. Each vertex corresponds to one observed spatial location, $\mathbf{s}$. An edge $(\mathbf{s}_p, \mathbf{s}_q) \in \mathcal{E}$ means that location $\mathbf{s}_p$ and $\mathbf{s}_q$ are neighbors. A weight function W assigns a weight to each edge in an edge set. Let $\boldsymbol{\theta}(\mathbf{s}) = \{\mu(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s})\}^T$, where $\mu(\mathbf{s}) = \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n)\}$, $\sigma(\mathbf{s}) = \{\sigma(\mathbf{s}_1), \ldots, \sigma(\mathbf{s}_n)\}$, and $\xi(\mathbf{s}) = \{\xi(\mathbf{s}_1), \ldots, \xi(\mathbf{s}_n)\}$. Then we can regulate $\mu(\mathbf{s})$, $\sigma(\mathbf{s})$, and $\xi(\mathbf{s})$ and estimate parameters by

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(\mathbf{s}) = \underset{\boldsymbol{\theta}(\mathbf{s})}{\operatorname{argmin}} & \sum_{i=1}^{n} -l\{\mu(\mathbf{s}_i), \sigma(\mathbf{s}_i), \xi(\mathbf{s}_i); y(\mathbf{s}_i)\} \\ & + \lambda_\mu \sum_{(\mathbf{s}_p, \mathbf{s}_q) \in \mathcal{E}} W_{p,q} \mid \mu(\mathbf{s}_p) - \mu(\mathbf{s}_q) \mid^k \\ & + \lambda_\sigma \sum_{(\mathbf{s}_p, \mathbf{s}_q) \in \mathcal{E}} W_{p,q} \mid \sigma(\mathbf{s}_p) - \sigma(\mathbf{s}_q) \mid^k \\ & + \lambda_\xi \sum_{(\mathbf{s}_p, \mathbf{s}_q) \in \mathcal{E}} W_{p,q} \mid \xi(\mathbf{s}_p) - \xi(\mathbf{s}_q) \mid^k, \end{aligned} \quad (7)$$

where $\lambda_\mu, \lambda_\sigma$, and λ_ξ are tuning parameters and $W_{p,q}$ is a weighted penalty factor that may depend on distances, for example. More details on the tuning parameter selection and weighted penalty factor are given in Section 3.3.

The fused penalty enforces spatial smoothness of parameters. Setting $k = 1$ represents the fused lasso penalty and creates sparsity in the difference between parameter estimates. Enforcing sparse differences results in clustered parameter estimates, meaning sites that are close in proximity may have identical parameter estimates. The fused lasso penalty has been used widely in regression models and an alternating direction method of multipliers (ADMM) algorithm was developed to estimate the parameters under such penalty. Setting $k = 2$ represents the fused ridge penalty and promotes similarity between parameter estimates. Unlike the fused lasso which

creates grouped estimates, the fused ridge produces smoothly varying parameter estimates.

3.2. Taylor Approximation for Optimization

The optimization in Equation (7) is difficult to solve because the likelihood is nonconvex and the parameters $\mu(\mathbf{s})$, $\sigma(\mathbf{s})$, and $\xi(\mathbf{s})$ are nonseparable in the likelihood. We are not aware of any solution that is guaranteed to find the global minimum. As a result we will use methods known to solve for the local minimum and it has been shown that these estimations are a good and efficient approximation (Tansey et al. 2018).

We begin by expanding $l\{\mu(\mathbf{s}_i), \sigma(\mathbf{s}_i), \xi(\mathbf{s}_i); y(\mathbf{s}_i)\}$ in a partial second-order Taylor approximation at the current iterate $\{z_{\mu(\mathbf{s}_i)}, z_{\sigma(\mathbf{s}_i)}, z_{\xi(\mathbf{s}_i)}\}$. Putting a full second-order Taylor approximation into quadratic form is extremely intricate, so we instead estimate each parameter sequentially allowing us to avoid having to compute cross derivatives in the Taylor series expansion. Consider, for example, the estimation of the location parameter $\mu(\mathbf{s})$, we assume $\sigma(\mathbf{s})$ and $\xi(\mathbf{s})$ are known and equal to the current iterate $z_{\sigma(\mathbf{s}_i)}$ and $z_{\xi(\mathbf{s}_i)}$, respectively. Then Equation (7) can be expressed solely in terms of $\mu(\mathbf{s})$ and the Taylor expansion turns Equation (7) into a fused regression problem. The estimation of $\mu(\mathbf{s})$, up to a constant term not depending on $\mu(\mathbf{s})$, becomes

$$\begin{aligned}\underset{\mu(\mathbf{s})}{\operatorname{argmin}} \quad & [\nabla l\{z_{\mu(\mathbf{s})}; y(\mathbf{s})\}]^T\{\mu(\mathbf{s}) - z_{\mu(\mathbf{s})}\} + \frac{1}{2}\{\mu(\mathbf{s}) - z_{\mu(\mathbf{s})}\}^T \\ & \times H\{z_{\mu(\mathbf{s})}; y(\mathbf{s})\}\{\mu(\mathbf{s}) - z_{\mu(\mathbf{s})}\} \\ & + \lambda_\mu \sum_{(\mathbf{s}_p,\mathbf{s}_q)\in\mathcal{E}} W_{p,q} \mid \mu(\mathbf{s}_p) - \mu(\mathbf{s}_q) \mid^k, \end{aligned} \tag{8}$$

where $\nabla l\{z_{\mu(\mathbf{s})}; y(\mathbf{s})\}$ and $H\{z_{\mu(\mathbf{s})}; y(\mathbf{s})\}$ are the gradient and Hessian with respect to the negative log-likelihood, $-l\{\mu(\mathbf{s}); y(\mathbf{s})\}$, evaluated at the current iterate $z_{\mu(\mathbf{s})}$. Since the true parameters are unknown, we will use the maximum likelihood estimates from the spatial GEV model to initialize the first iteration.

Ignoring terms that are constant in $\mu(\mathbf{s})$, the solution of Equation (8) can be expressed as

$$\begin{aligned}\widehat{\mu}(\mathbf{s}) = \underset{\mu(\mathbf{s})}{\operatorname{argmin}} & \sum_{i=1}^{n} \frac{\{y_i - x_i\mu(\mathbf{s}_i)\}^2}{2} \\ & + \lambda_\mu \sum_{(\mathbf{s}_p,\mathbf{s}_q)\in\mathcal{E}} W_{p,q} \mid \mu(\mathbf{s}_p) - \mu(\mathbf{s}_q) \mid^k, \end{aligned}$$

where

$$\begin{aligned} y_i &= \left[z_{\mu(\mathbf{s}_i)} - \frac{\frac{\partial l\{\mu(\mathbf{s}_i); y(\mathbf{s}_i)\}}{\partial \mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}{\frac{\partial^2 l\{\mu(\mathbf{s}_i); y(\mathbf{s}_i)\}}{\partial \mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}} \right] \\ &\quad \times \sqrt{\frac{\partial^2 l\{\mu(\mathbf{s}_i); y(\mathbf{s}_i)\}}{\partial \mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}} \\ x_i &= \sqrt{\frac{\partial^2 l\{\mu(\mathbf{s}_i); y(\mathbf{s}_i)\}}{\partial \mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}. \end{aligned}$$

Replacing $\mu(\mathbf{s})$ with $\sigma(\mathbf{s})$ or $\xi(\mathbf{s})$ will provide estimates for the scale or shape parameter, respectively, the details and derivatives of which are shown in Appendix B.

When the support of the density depends on unknown parameters, as in Equations (4) and (6), MLEs do not satisfy the classical regularity conditions. See Smith (1985) for more details. It is known that the maximum likelihood estimation of multi-parameter distributions can break down because the likelihood is unbounded (Cheng and Amin 1983), which can lead to issues of parameter divergence. To mitigate rare instances of divergence in parameter estimation in the proposed models, conservative bounds can be specified and if a parameter is estimated outside those bounds it will be re-initialized and re-estimated.

3.3. Implementation of Fused Models

To select the set of edges, $\mathcal{E}$, we use a minimum spanning tree (MST) to define neighboring sites (Li and Sang 2019). A MST is a graph that connects all of the sites together, without any cycles and with the minimum possible total edge length. Using a MST removes redundant pairs and allows for faster computation when working with large datasets. We then apply an inverse distance penalty factor $W_{p,q} = 1/||s_p - s_q||$ to the neighboring sites. This enforces more similarity between sites that are closer in distance and allows flexibility when the sites are farther apart. The choice of $W_{p,q}$ is not unique and there is no known optimality on how to make a choice. Since the strength of dependency of spatial data often depends on distance, we simply chose the inverse distance weighting that has been one of the most popular choice for interpolating geographic data (Burrough and McDonnell 1998; Longley et al. 2001). The connecting edges can be written as a matrix D, that is of dimension $(n - 1) \times n$. Without loss of generality, we add a row to D but set its corresponding penalty factor equal to zero simply for the purpose of transforming D to full rank. This allows for an easy bijective transformation from the vector of parameters $\boldsymbol{\theta}(\mathbf{s})$ into a vector of parameter differences such that the fused models can be implemented using the "glmnet" package in R.

The choice of λ in the fused lasso and fused ridge penalties is important because too small of a λ will result in overfitting and too large of a λ will result in oversmoothing and loss of spatial structure. In practice, the optimal λ can be determined via some data-dependent model selection criteria (Li and Sang 2019), such as generalized cross-validation (Golub, Heath, and Wahba 1979), Akaike information criterion (AIC), Bayesian information criterion (BIC) (Schwarz 1978), and the composite likelihood information criterion (CLIC) (Varin and Vidoni 2005). We fit our model using the "glmnet" package in R which uses the deviance for the cross-validation metric. For a Gaussian model, this deviance is the sum of squared errors (SSE).

The GEV parameters, $\{\mu(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s})\}$, can be estimated either sequentially or independently. Sequential parameter estimation means that the first parameter is initialized using the spatial GEV and the remaining parameter estimations are initialized with the prior parameter iteration results. There are six different combinations for estimating sequentially: $\{\mu \rightarrow$

$\sigma \rightarrow \xi; \mu \rightarrow \xi \rightarrow \sigma; \xi \rightarrow \mu \rightarrow \sigma; \xi \rightarrow \sigma \rightarrow \mu; \sigma \rightarrow \xi \rightarrow \mu; \sigma \rightarrow \mu \rightarrow \xi\}$ with $(\mathbf{s})$ suppressed from the notation. Independent parameter estimation means that the estimation of each parameter is initialized using the spatial GEV, and not reliant on the other parameter estimation results. In order to demonstrate the difference between these two estimation schemes, we compared parameter estimates computed both independently and sequentially through simulations in Appendix E. The simulation results in Appendix E showed that the sequential estimation results were best when $\sigma(\mathbf{s})$ is estimated prior to $\xi(\mathbf{s})$. The estimation of $\mu(\mathbf{s})$ has minimal impact on the estimation of the other parameters. Because the accuracy of the Taylor approximation is dependent on the values chosen to expand about, the closer the value is to the true value the more accurate the results, it is expected that the average root mean squared error (RMSE) for the parameters would increase when estimated independently compared to being estimated sequentially. Following this notion, for the simulation study in Section 4 we choose to estimate the parameters in the order of $\sigma(\mathbf{s}) \rightarrow \xi(\mathbf{s}) \rightarrow \mu(\mathbf{s})$, because $\sigma(\mathbf{s})$ shows the greatest improvement over the Spatial GEV allowing for better initialization when estimating $\xi(\mathbf{s})$, and so forth. However, the results in Appendix E also show that for the fused ridge model estimating parameters independently is still an improvement over the spatial GEV model, and only a relatively small change compared to the sequential estimation making it a reasonable alternative in applications where sequential estimation is not practical.

3.4. Bootstrap to Quantify Uncertainty

The delta method, profile likelihood, and bootstrap are common methods to estimate the CI for return levels. The delta method (Coles 2001) is a classical means to compute CI by assuming the asymptotic normality of return level estimates. The profile likelihood is computationally efficient, however, it is only available for univariate GEV and may fail to keep the spatial smoothness of spatial extremes. For return levels of spatial extremes, the bootstrap (Davison and Hinkley 1997) is considered a popular choice to compute their CIs. The accuracy of a parametric bootstrap CI depends on how well the fitted parametric model approximates the true underlying model. The nonparametric block bootstrap has the advantage of preserving the spatial dependence structure of extremes by bootstrapping only in the time dimension. If the fitted model cannot approximate the true dependence structure, then the nonparametric block bootstrap is preferred (Cao and Li 2018).

We will use block bootstrap to quantify uncertainty. This is performed by subsampling spatial "blocks" in time. The return level estimate from each bootstrapped data constitutes a random sample drawn from the distribution of the return level estimator, and thus, the α/2th and (1-α/2)th quantiles of the bootstrap estimates offer a bootstrap CI at the significance level α. The uncertainty estimation for our simulation studies are shown in Appendix G.

4. Monte Carlo Studies for Model Comparison

We conduct simulation studies using both stationary and nonstationary data to assess the performance of parameter and return level estimation using our fused spatial GEV and spatial GPD compared to the five methods introduced in Section 2 and Appendix A.

4.1. Simulation Settings

We first simulate data from a stationary max-stable field using the Schlather model with marginal GEV parameters mimicking estimates from observed annual maximum daily precipitation across Colorado's Front Range (Tye and Cooley 2015). Spatial locations, $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, are randomly generated from the uniform distribution over a $[0, 20] \times [0, 20]$ spatial domain, $\mathcal{D}$. The number of spatial locations is chosen to be $n = 200$, and the number of replicates is set to be $m = 50$ for the spatial GEV and $m = 500$ for the spatial GPD. We spatially smooth the 90th percentile at each site using a Gaussian kernel to select the threshold, and then collect exceedances when fitting the GPD.

Spatially varying GEV parameters, $\boldsymbol{\theta}(\mathbf{s}) = \{\mu(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s})\}^T$, are sampled from stationary and isotropic Gaussian processes $GP\{m(\mathbf{s}), \Sigma\}$ with mean functions $m(\mathbf{s}) : \mathcal{D} \rightarrow \mathbb{R}$ and powered exponential covariances $\Sigma : \mathcal{D}^2 \rightarrow [0, \infty)$. The parameter choices follow Tye and Cooley (2015). Shape parameters are resampled until $\xi(\mathbf{s}) > 0$ for all $\mathbf{s}$, to ensure data are heavy-tailed (Hewitt et al. 2019). Further details of the model generating configurations are specified in Table C1 of Appendix C. We use the generated GEV parameters to transform the marginal distribution of the max-stable process from a unit Fréchet distribution to a more general GEV distribution.

Nonstationary data will be generated by deforming locations of a stationary max-stable field. Spatial locations, $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, are again randomly generated from the uniform distribution but now over a $[0, 1] \times [0, 1]$ spatial domain. Once the stationary data is generated we transform it into nonstationary data by squaring the spatial locations, $\{\mathbf{s}_1^2, \ldots, \mathbf{s}_n^2\}$ and treating the squared locations as the actual locations for simulated data. This creates stronger dependence in the north and east than in the south and west. The domain is set to $[0, 1] \times [0, 1]$ to ensure the domain of the warped locations remains the same. Correspondingly, in the data generation the scale parameters are also scaled by a factor of 20 to provide equivalent analysis, as specified in Table C2 of Appendix C.

4.2. Spatial Extremes With GEV Marginals

We compare our fused spatial GEV models under either ridge penalty or lasso penalty to the spatial GEV, Schlather, Bayesian HKEVP, and latent variable model (LVM). For the spatial GEV, Schlather, and LVM Gaussian process we model their marginal parameters as in Equation (5) with covariates $X_\mu(\mathbf{s}) = (1, s_x)$ and $X_\sigma(\mathbf{s}) = (1, s_y)$, where s_x and s_y represent the longitude and latitude coordinates, respectively. The Schlather model uses pairwise likelihoods. For Bayesian HKEVP, we place knots at each location of observations. The parameters of the fused spatial models were initialized at the spatial GEV estimates and estimated with four iterations. The number of iterations was chosen through trial and error. For each method we run the simulation 200 times and use the RMSE of parameter and return level estimation to compare performance.

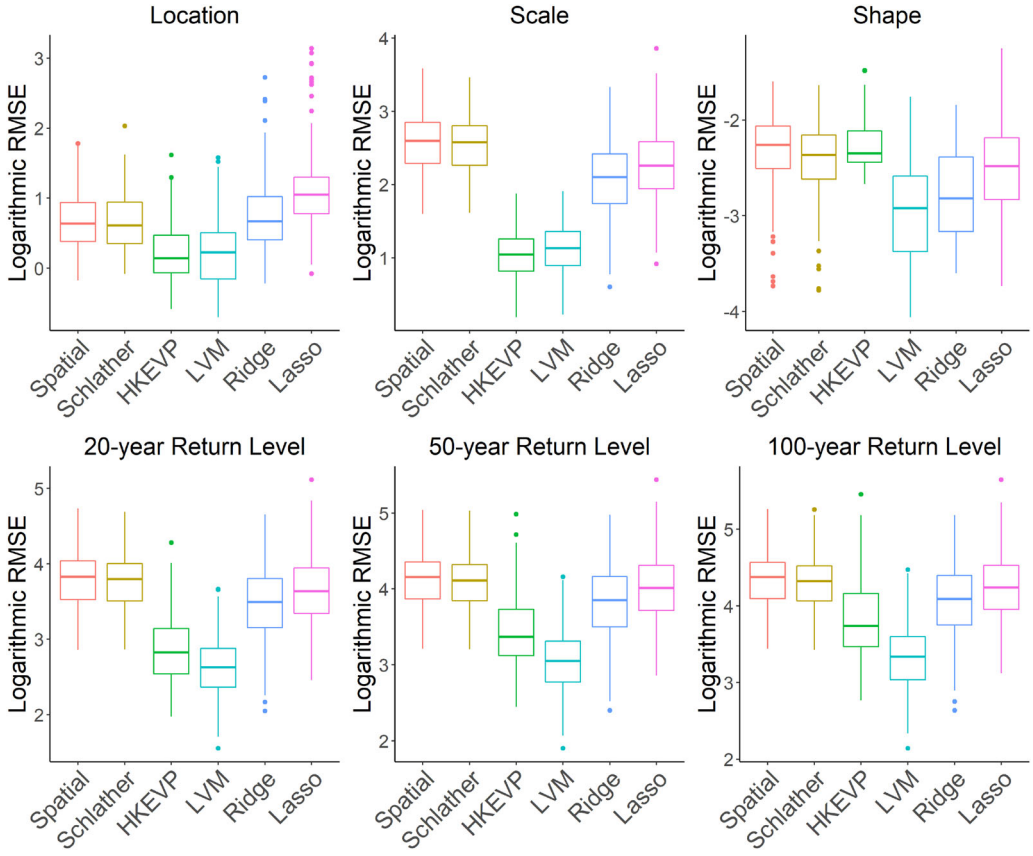

Figure 1. Boxplots of logarithmic RMSE for the GEV parameters and return level estimates with stationary spatial extremes and six models: Spatial GEV, Schlather, Bayesian HKEVP, Latent Variable, Fused spatial GEV under Ridge and Lasso penalty.

4.2.1. Stationary Spatial Extremes

We simulate stationary spatial extremes as specified in Section 4.1. The LVM and HKEVP model runs 7500 iterations with a 5000 burn-in. Figure 1 shows the logarithmic RMSE for the parameter and return level estimates of the 200 simulations. A summary of the average RMSE for the parameter and return level estimates of the six models is provided in Appendix D. It is seen from Figure 1 that overall the LVM and HKEVP models perform best at the expense of a larger computation time of on average 345.15 and 15,694 sec, respectively. In contrast, the fused ridge and fused lasso computation times were on average 4.32 sec and 1.90 sec, respectively. The spatial GEV took on average 1.79 sec and the Schlather model took on average 61.11 sec. Latent variable models are known to model the marginal distribution of spatial extremes well, so it is not surprising that this model performed best (Davison, Padoan, and Ribatet 2012). However, unpublished simulation results showed that the advantage of the LVM over the fused models diminishes as the spatial extremes process becomes smoother. Because the HKEVP model uses Gaussian processes as priors for the parameters which is conducive to the simulated parameters, it is not surprising that this model performs well.

The shape parameter which is generally difficult to estimate, is captured best by the LVM and fused ridge model. The shape parameter estimation has an increasing impact on return level as the return period increases. As a result, the gap between the HKEVP and fused ridge model decreases as the return level period increases. While the fused lasso model performs relatively similar to the spatial GEV and Schlather model, the fused ridge model outperforms both the spatial GEV and Schlather in every aspect except for the location parameter. This is likely because the simulated parameters are smoothly varying favoring the fused ridge, which provides smoothly varying estimates, over the fused lasso, which provides clustered estimates. The improvement in return level estimation of the fused ridge and fused lasso leads them to be appealing options for modeling large datasets when computational efficiency is desired.

Table 1. The average computation time in seconds of 10 simulations for the Latent Variable, Fused ridge, and Fused lasso GEV models.

	LVM	Fused ridge	Fused lasso
200 Sites	345.15	4.32	1.90
400 Sites	2,362.70	17.00	6.83
800 Sites	18,933.80	89.26	36.67

While the latent variable model performs the best overall, it is not always practical for large datasets due to the computational burden. To demonstrate how the computation time increases as the number of sites considered increases, we conduct 10 sample simulations of 200, 400, and 800 sites and record the average computation time. The computation time of the LVM is only for 7500 iterations and a 5000 burn-in. Further studies would need to be conducted to ensure convergence in real applications. Table 1 shows the computational advantage our model has in large datasets.

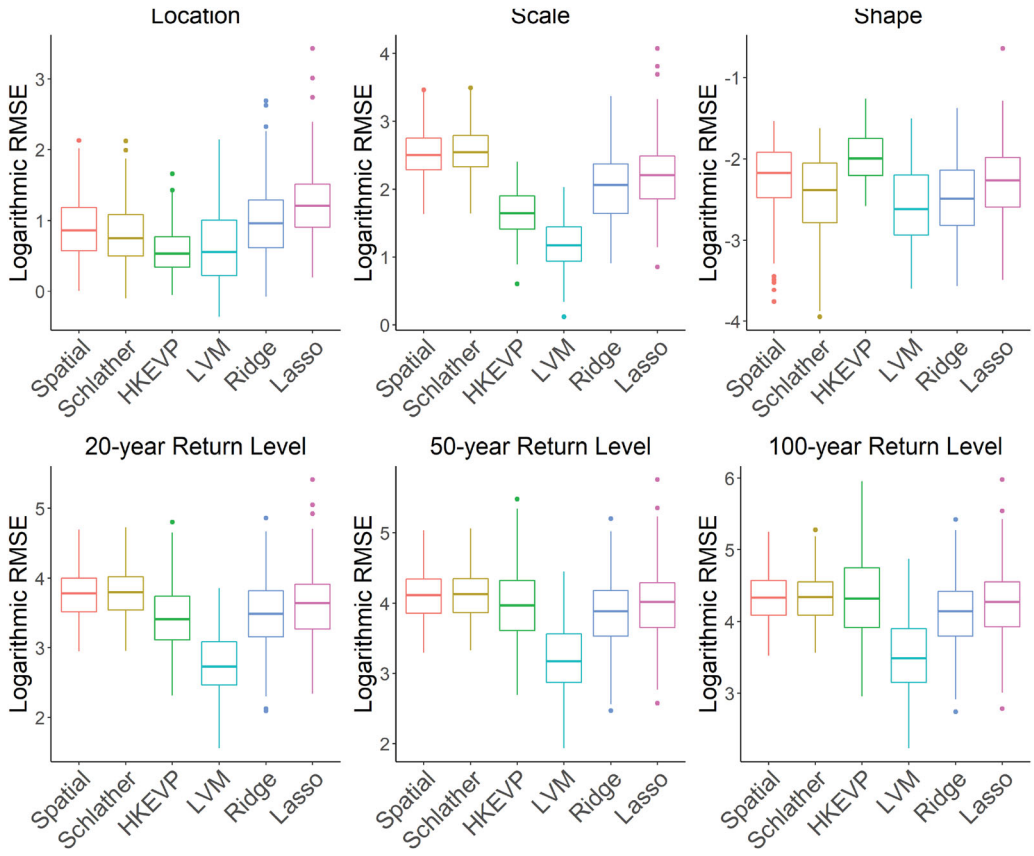

Figure 2. Boxplots of the logarithmic RMSE for the GEV parameters and return level estimates with nonstationary spatial extremes and six models: Spatial GEV, Schlather, Bayesian HKEVP, Latent Variable, Fused spatial GEV under Ridge and Lasso penalty.

A caveat in fitting the fused ridge and fused lasso model should be mentioned. If the second derivative of a site in the Taylor approximation is negative, then that site cannot be used due to being incompatible with the fused regression. The number of site failures is dependent on the initial parameters used in the Taylor approximation. Therefore, initial parameters should be chosen such that a majority, if not all, of the sites are retained. From experience, this can generally be accomplished by setting the initial location parameters to be a constant minimum or maximum of the data.

4.2.2. Nonstationary Spatial Extremes

The simulation study in Section 4.2.1 is based on stationary data with parameters from a Gaussian process. However, in reality, not all spatial extremes follow this ideal data structure. This section compares the model performance with nonstationary data that is generated by deforming locations of a stationary max-stable field. The LVM and HKEVP model runs 7500 iterations with a 5000 burn-in.

Figure 2 provides boxplots of the logarithmic RMSE for the parameter and return level estimates. The corresponding averages are provided in Table D2 of Appendix D. With the nonstationary data simulated in this section, the LVM and fused ridge model have the best shape parameter and return level estimates. The LVM is again superior in terms of return level estimation. The HKEVP model performs similarly to the fused ridge in the 20-year return level, however the HKEVP model performs significantly worse as the return level period increases due to its larger error in the shape parameter estimation. Perhaps even more importantly, the fused ridge has an average computation time of 4.79 sec compared to the LVM and HKEVP models which have average computation times of 346.98 and 15,551 sec, respectively. The spatial GEV, Schlather, and fused lasso model all perform similarly with average computation times of 1.97, 77.50, and 2.15 sec, respectively.

4.3. Spatial Extremes With GPD Marginals

For peaks over threshold, we again compare the model performance with both stationary and nonstationary data. The basic settings for locations and replicates of the simulated data are already described in Section 4.1. To calculate the threshold for each site, we first choose the 90th quantile at each site and then spatially smooth the values using a Gaussian kernel known as the Nadaraya-Watson smoother (Nadaraya 1964, 1989; Watson 1964). In practice, choosing a threshold is a delicate procedure weighing the bias versus precision tradeoff and is usually done using diagnostic plots showing how quantities, such as the mean residual life, vary as the threshold changes (Coles 2001; Cooley, Nychka, and Naveau 2007). The shape parameter of the GPD remains the same as in the GEV, and the true scale parameter of the GPD can be calculated based on its relationship with the parameters in the GEV and threshold as described right below (2). We compare our fused spatial GPD with ridge or lasso penalty only to the spatial GPD and LVM, as the Bayesian HKEVP did not address the situation with GPD marginals.

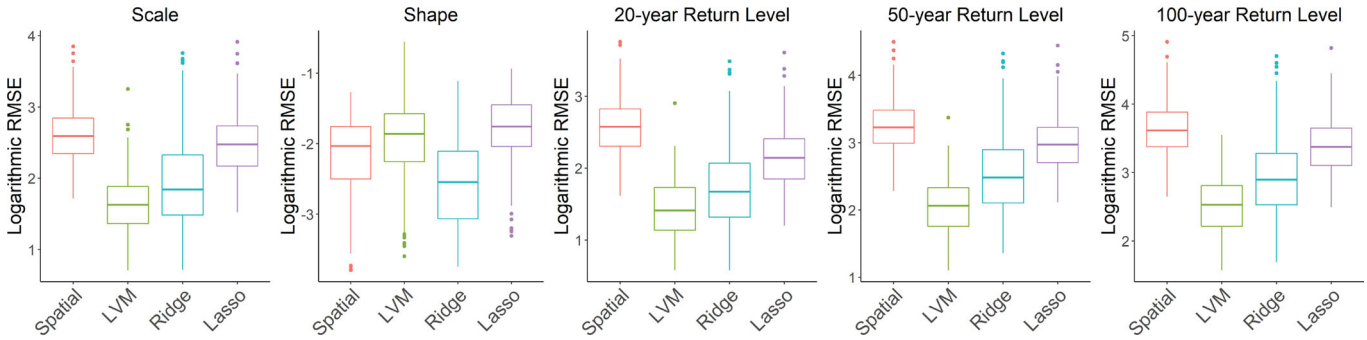

Figure 3. Boxplots of the logarithmic RMSE for the GPD parameters and return level estimates with stationary spatial extremes and four models: Spatial GPD, Latent Variable, Fused spatial GPD under Ridge and Lasso penalty.

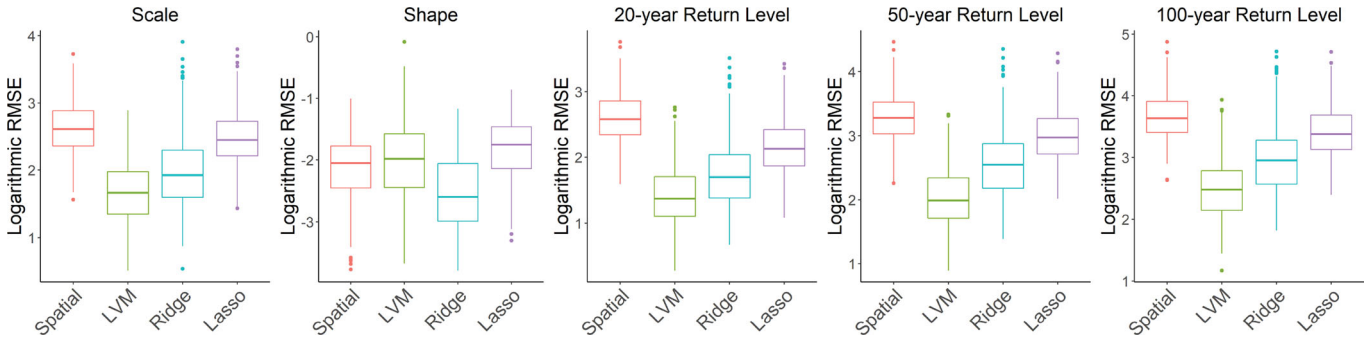

Figure 4. Boxplots of the logarithmic RMSE for the GPD parameters and return level estimates in the nonstationary spatial extremes and four models: Spatial GPD, Latent Variable, Fused spatial GPD under Ridge and Lasso penalty.

For spatial GPD, the scale parameters are modeled as $\tilde{\sigma}(\mathbf{s}) = \boldsymbol{\beta}_{\tilde{\sigma}}^T X_{\tilde{\sigma}}(\mathbf{s})$, where $X_{\tilde{\sigma}}(\mathbf{s}) = (1, s_x, s_y)$ and s_x and s_y represent the longitude and latitude coordinates, respectively.

The parameters of the fused models were initialized at the spatial GPD estimates and estimated with six iterations for the fused ridge and one iteration for the fused lasso. The number of iterations was chosen through trial and error. Similar to the simulations with block maxima, when some locations have a negative second derivative in the Taylor approximation, the choice of initial values should be tweaked such that a majority, if not all, of the sites are retained. The number of iterations was chosen through trial and error. For each method, we run the simulation 200 times and use the RMSE of return level estimation to compare performance.

4.3.1. Stationary Data

Figure 3 provides boxplots of logarithmic RMSEs for the parameter and return level estimates of the four models. The corresponding averages and computation times are provided in Table D3 of Appendix D. The fused ridge and fused lasso models perform similarly and both outperform the spatial GPD in the scale parameter and return level estimation. The RMSE of the shape parameters with both fused models is higher than with the spatial GPD. Further investigation shows that the fused models actually capture the spatial variability of the shape parameters well; however the fused models tend to overestimate the shape parameters while the spatial GPD fails to capture the variation in the shape parameters. While the fused models outperform the LVM in terms of the shape parameter estimation, the LVM performs better in terms of return level estimation. The LVM was run with 15,000 iterations and a 5000 burn-in. LVMs are known to capture the marginal distribution well. However, the computation time was on average 635.36 sec compared to the fused ridge and fused lasso computation times of 13.80 and 0.86 Sec, respectively.

4.3.2. Nonstationary Data

Figure 4 summarizes the parameter and return level estimates of the four models using 200 simulations. Table D4 in Appendix D provides the corresponding average RMSE and computation times. The results for the nonstationary data are very similar to that of the stationary data in Section 4.3.1. The fused ridge has an average computation time of 14.49 sec compared to the LVM model which has an average computation times of 620.83 sec.

5. Application to Spatial Weather Extremes

5.1. Return Level Estimation for Precipitation Data

We analyze climate model output on annual maximum daily precipitation over historical (1969–2000) and future conditions (2039–2070) for 2622 sites across the continental United States. These data were provided by the North American Regional Climate Change Assessment Program (NARCCAP) and can be downloaded from the website *http://www.narccap.ucar.edu/index.html*. The NARCCAP chose the A2 emissions scenario, which is described in Nakicenvoic et al. (2000). The output data were produced using the Geophysical Fluid Dynamics Laboratory's AM2.1 climate model with 50km resolution. NARCCAP provides eight 3-hr precipitation rates each day, and we compute the daily total by summing these eight values and multiplying by three. We then analyze the annual maximum of the daily

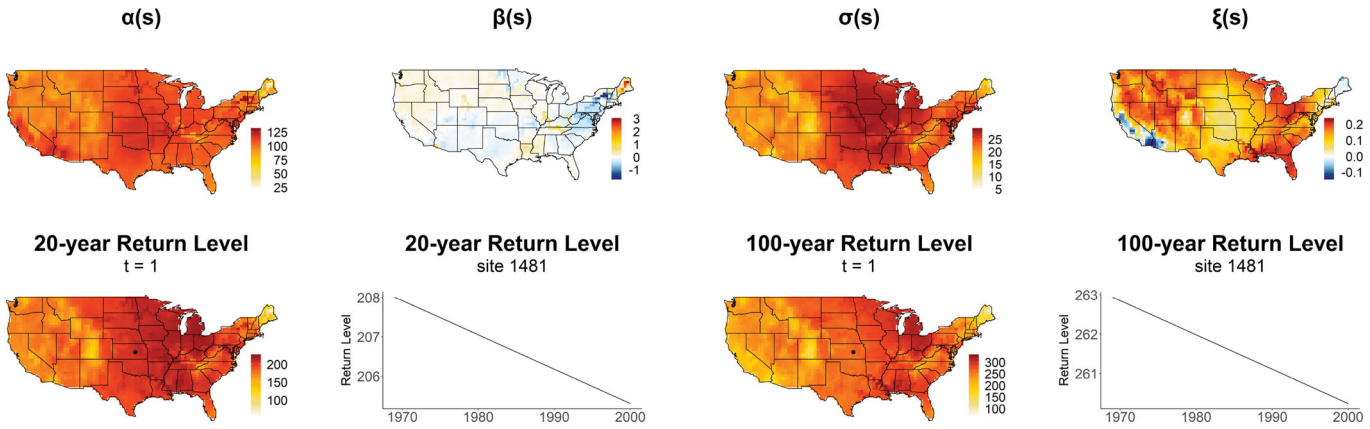

Figure 5. Estimated GEV parameters, 20-, and 100-year return level for the historical simulation using the fused ridge model. The spatially varying return levels are shown at time t=1 (year 1969) and the time varying return levels are shown for site 1481, located at the black dot. All units are mm/h.

precipitation totals from the two time-slices separately using the fused spatial GEV model under either ridge or lasso penalty.

In the parameter initialization, we basically follow the estimates from the spatial GEV but tweak them slightly to ensure a majority of the sites are retained in the estimation. The nature of the fused penalty in our fused spatial GEV restricts the parameter estimates and thus the return level estimates to only site-wise locations. In order to estimate the return level at unknown locations, we interpolate using ordinary kriging. Kriging is a traditional approach for spatial data interpolation that exploits the spatial correlation between observations. Cao and Li (2018) showed that kriging interpolated return levels performed qualitatively similar as the return level estimation based on max-stable models when data follow a max-stable model.

Given the long time period considered, it is unlikely that the data is stationary across time. To account for the non-stationarity, we incorporate a simple linear time trend into the location parameter. We model $\mu(\mathbf{s}, t) = \alpha(\mathbf{s}) + \beta(\mathbf{s})t$, where $\alpha(\mathbf{s})$ is the spatially varying location parameter, $\beta(\mathbf{s})$ is the spatially varying coefficient associated with the time trend, and $t = 1$ for years 1969 and 2039, up to $t = 32$ for years 2000 and 2070. Then in Equation (3), we replace $\mu(\mathbf{s})$ with $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$, and optimize these parameters sequentially. The computation time for the fused ridge was 11,801 seconds for the historical simulation and 13,882 seconds for the future simulation. The computation time is based on an i7 core processor with 4 Gb of RAM.

Figure 5 shows the estimated GEV parameters, 20-, and 100-year return level for the historical simulation using the fused ridge model. Since we model the location parameter as a function of time, the return level is also a function of time. For illustration purposes, we plot the 20- and 100-year return level map at time, $t = 1$, and the time varying return level for a single site, $s = 1481$. The results for the fused lasso model are very close to those of the fused ridge so we omit the fused lasso plots. The estimated location and scale parameters are highest in the midwest. There is a positive trend in the northwest and the largest negative trend is in Virginia. The generally positive shape parameter indicates a heavy tailed distribution with no upper bound. The 20-year return level tends to mimic the high and low estimates of the location and scale parameters. The 100-year return level shows an increased precipitation in Ohio and Kentucky heavily influenced by the larger shape parameter. However, the negative trend in Ohio and Kentucky suggests the return level will decrease when estimated at a later time, for example at $t = 32$. The corresponding standard errors of the estimates in Figure 5 using 200 bootstrap samples are provided in Figure F1 of Appendix F. We use the nonparametric block bootstrap to quantify uncertainty.

The estimated change in the GEV parameters, 20-, and 100-year return level estimates from the historical simulation to the future conditions simulation for the fused ridge model are shown in Appendix F along with the corresponding block bootstrap standard errors. Again, the plots for the fused lasso model are very similar so we omit the results. In the future conditions simulation there is a relatively small increase in all parameter estimations, except the trend. The largest location and scale changes are in the northeast. The shape parameter shows significant increases outside of the southeast. These increases were most impactful on the return levels in the midwest and northeast.

A subset of this data was modeled in Reich and Shaby (2012) using the HKEVP, which included the 697 stations east of −90 degrees longitude. Our fused model results are similar to the HKEVP model for the historical simulation in that the shape parameter is generally positive and largest in Florida, however our shape parameters are more moderate. The estimated scale parameters have a similar range, the main difference being the HKEVP estimates the largest scale parameters in the southeast where our fused model estimates the largest scale parameters in the midwest. The estimated return levels are very comparable in the north. In the south our estimates are less volatile than the HKEVP model, which can be attributed to the larger scale and shape parameters in the HKEVP model. When comparing the historical and future scenarios both our fused model and the HKEVP model estimates an increase in return level for the majority of the spatial domain.

5.2. Annual Maximum Temperature Change

To demonstrate the ability of the fused models to handle even larger datasets, we analyze annual maximum temperature data for 8,125 sites across the continental US for the years 1898–

1997. The data were provided by the National Corporation for Atmospheric Research (NCAR) and is available upon request from Dr. Douglas Nychka. Observed data at each station was used when available and any missing station values were filled in using spatial statistics to produce a complete dataset (Johns et al. 2003). A more detailed description of the data can be found at *https://www.image.ucar.edu/Data/US.monthly.met/*. We split the data into two equal time periods of 50 years and estimate the return levels for each time period separately.

In the parameter initialization, we follow the estimates from the spatial GEV with minor adjustments to ensure a majority of the sites are retained in the estimation. Estimates are interpolated over the whole spatial domain using ordinary kriging and standard errors are computed using 200 nonparametric block bootstrap samples. The standard error plots are available in Appendix F. Given the long time period, we model the location parameter as in Section 5.1 to account for nonstationarity, where $t = 1$ represents years 1898 and 1948, up to $t = 50$ for years 1947 and 1997. The results shown are for the fused ridge model, the fused lasso model had very similar results so we omit the fused lasso plots. The computation time for the fused ridge was approximately 13.53 hr for each simulation. This computation time could be further reduced by one third by estimating the parameters independently and in parallel for comparable results.

The return level estimates in the top row of Figure 6 suggest the hottest maxima temperatures are in the southwest while the lowest maxima temperatures are along the rocky mountains in Colorado and Wyoming. Lower return levels are also expected in the northeast and along the Washington coastline. This follows the intuition of what one generally expects when thinking of historical temperatures across the United States.

It is known that average temperatures in the United States since 1900 have warmed on average, however it is less clear whether maximum temperatures have changed during this period (Lee, Li, and Lund 2014). The bottom row of Figure 6 shows the change in return level estimates between the two time periods at time t=1, 1898 and 1948. From the plots we can see the temperature has the largest increases in the midwest. At time $t = 1$, 3,178 of the 8125 stations had a decrease in annual maximum temperature for the 100-year return level. Compared to time $t = 50$, 6264 of the 8125 stations had a decrease in annual maximum temperature. The GEV parameter plots are provided in Appendix F to show the predominantly negative time trend. However, none of the changes are statistically significant at the 95% confidence level. The decreases may seem surprising, however a 'warming hole' in the eastern United States has been previously noted (Lund, Seymour, and Kafadar 2001). Lee et al. (2014) also found when analyzing maximum temperature trends over the years 1890–2010 across the United States that 583 of 923 stations had negative trends.

6. Discussion

In order to improve the return level estimation for large spatial extremes datasets, we propose fused spatial GEV and fused spatial GPD models with varying coefficients under either a ridge or lasso penalty. Our proposed models are flexible in parameterization and thus are able to capture the spatial variability of the data better than the spatial GEV and spatial GPD with parametric parameter specifications. Our models require no assumption of stationarity and are significantly more computationally efficient compared to Bayesian models. The simulation study showed that the proposed models outperform spatial GEV, spatial GPD, and max-stable models when the marginal GEV and GPD vary spatially. Our models also outperform Bayesian models when the spatial extremes process is smooth or when the spatial extremes data show nonstationary dependence. While in general Bayesian models yield the most accurate return level estimation, their extensive computation may discourage users. The nonrestrictive parameter estimates from the proposed fused models could be used as initial values to accelerate the convergence when fitting a complicated and time-consuming Bayesian model. The fused spatial models with ridge penalty are obviously suitable for smoothly varying marginal behavior, while if the marginal GEV or GPD distribution of the spatial extremes is spatially clustered, then the fused lasso models are expected to be more appropriate. The data applications demonstrate the ability of the fused spatial extremes models to produce practical return level maps for large datasets.

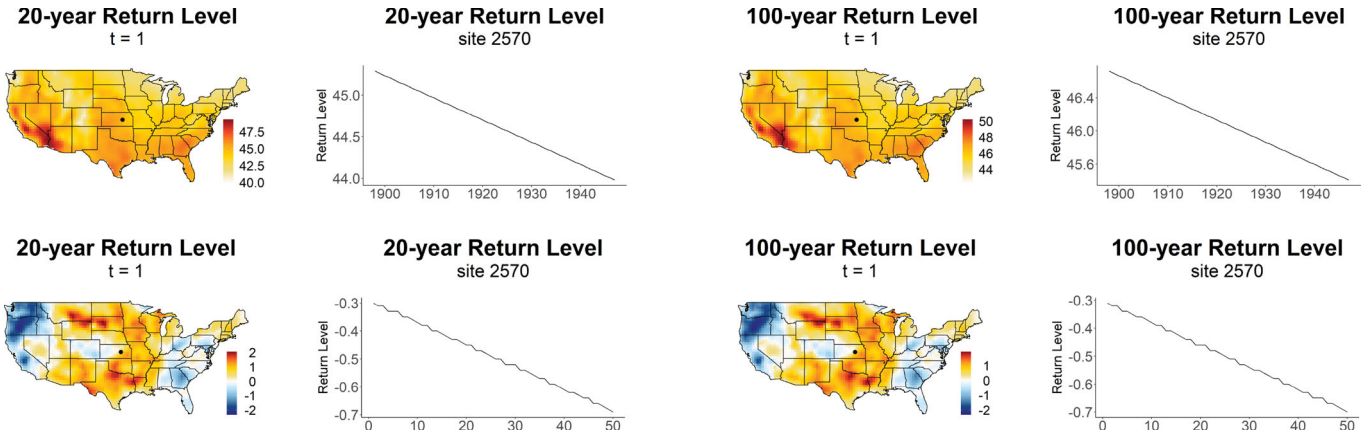

Figure 6. Top: estimated 20- and 100-year return level for the years 1898–1947 using the fused ridge model. Bottom: estimated change in 20- and 100-year return level between the years 1898–1947 and 1948–1997 using the fused ridge model. The spatially varying return levels are shown at time $t = 1$ (1898 and 1948) and the time varying return levels are shown for site 2570, located at the black dot. All units are in degrees Celsius.

The undirected edges used to define neighboring sites could be determined by other methods such as king neighbors for gridded data rather than a minimum spanning tree, but additional work may be solicited to remove the redundant edges to allow for invertibility of D. One restriction of the fused models is that the model fitting only estimates parameters and return levels at observed locations but not in a continuous domain. We addressed this issue by interpolating estimated parameters over the whole domain using kriging in the data application. Another limitation of the fused models lies in its implementation being dependent on the initialized values. It could be of interest to further explore the robustness of the initialized values and consistency of the estimation.

Supplementary Material

Appendices: Appendix A provides review of a max-stable model, hierarchical kernel Bayesian model, and a latent variable model. Appendix B contains the derivation of the Taylor approximation as well as the first and second derivatives of the GEV and GPD distributions. Appendix C contains model generating configurations for the simulation studies. Appendix D contains tables of the average RMSE and computation times for the simulation studies conducted in Section 4. Appendix E contains results of an additional simulation study comparing parameter estimation procedure. Appendix F contains additional results for the data applications in Section 5. Appendix G contains the results of the bootstrap simulation study.
R Code: This file contains the code used for Sections 4 and 5. A README file describes the contents.

Acknowledgments

The authors thank the editors and the two anonymous reviewers for their constructive suggestions that have greatly improved the content and presentation of this article. We also thank Dr. Dan Cooley for sharing his computing code.

Funding

Li acknowledges partial support from the NSF grants AGS-1602845 and DMS-1830312. Reich acknowledges partial support from The King Abdullah University of Science and Technology (grant no. 3800.2) and National Institutes of Health (NIH grant no. R01ES027892).

Appendix A

A.1. Schlather Model

A max-stable model has the GEV as its marginal distribution while accounting for the spatial dependence. We will use the Schlather (2002) model as an example to illustrate max-stable models because it is the model we chose to generate simulated data from in Section 4.

Let $Z(\mathbf{s})$ be a max-stable process, then due to de Haan (1984) any max-stable process can be described through the spectral representation, that is, $Z(\mathbf{s}) = \max_{j\geq 1}\zeta_j W_j(\mathbf{s})$, where ζ_j are points of a Poisson process on $(0,\infty)$ with intensity measure $d\zeta/\zeta^2$ and $\{W_j\}$ are independent copies of a stochastic process W with continuous sample paths satisfying $E[W(\mathbf{s})_+] = 1$. Different choices of the spectral process $W_j(\mathbf{s})$ leads to different max-stable models. In the Schlather model, $W_j(\mathbf{s})$ are stationary Gaussian processes with correlation function $\rho(\mathbf{s})$, scaled such that $E[W(\mathbf{s})_+] = 1$. One example for correlation function is the powered exponential function $\rho(h) = (1-\eta)\exp\{-(h/\lambda)^\nu\}$, where $\eta \in [0,1), \nu \in (0,2]$, and $\lambda > 0$ are the nugget, smooth, and range parameter, respectively, which need to be estimated (Sebille, Fougeres, and Mercadier 2016).

Max-stable models generally rely on pairwise likelihood functions, since the joint likelihood of spatial extremes is typically not available. The bivariate cdf of $Z(\mathbf{s})$ at a pair of sites $\mathbf{s}_i$ and $\mathbf{s}_j$ is given by $P(Z(\mathbf{s}_i) \leq z_i, Z(\mathbf{s}_j) \leq z_j) = \exp\{-V(z_i,z_j)\}$, where $V(z_i,z_j) = \frac{1}{2}(\frac{1}{z_i} + \frac{1}{z_j})\left(1+\sqrt{1-2\frac{\{\rho(h)+1\}z_i z_j}{(z_i+z_j)^2}}\right)$ is the exponent function that quantifies the spatial dependency, and h is the Euclidean distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$.

A.2. Bayesian Model

We use the hierarchical kernel extreme value process (HKEVP) proposed by Reich and Shaby (2012) as an example to illustrate Bayesian models. Suppose $y(\mathbf{s}) \sim \text{GEV}(\mu(\mathbf{s}),\sigma(\mathbf{s}),\xi(\mathbf{s}))$ and $Z(\mathbf{s}) \sim \text{GEV}(1,1,1)$ is the associated simple max-stable process. The spatial process $Z(\mathbf{s})$ can be constructed as the product of two independent processes, $Z(\mathbf{s}) = U(\mathbf{s})\theta(\mathbf{s})$. The nugget, $U(\mathbf{s})$, is spatially independent $\text{GEV}(1,\alpha,\alpha)$ with the parameter $\alpha \in (0,1)$ controlling the relative contribution of the nugget effect. The residual spatial process $\theta(\mathbf{s})$ is defined based on a linear combination of L kernel basis functions $w_l(\mathbf{s}) \geq 0$ satisfying $\sum_{l=1}^L w_l(\mathbf{s}) = 1$. Specifically, $\theta(\mathbf{s}) = \{\sum_{l=1}^L A_l w_l(\mathbf{s})^{1/\alpha}\}^\alpha$ where A_i are the basis function coefficients satisfying $A_l \overset{\text{iid}}{\sim} \text{PS}(\alpha)$, a positive stable distribution with characteristic exponent α, ensuring max-stability and unit Fréchet marginals of $Z(\mathbf{s})$.

The hierarchical model is given by

$$Y(\mathbf{s}_i)|A_1,...,A_L,\mu,\sigma,\xi,\alpha \sim \text{GEV}\{\mu^*(\mathbf{s}_i),\sigma^*(\mathbf{s}_i),\xi^*(\mathbf{s}_i)\}$$
$$A_l \overset{\text{iid}}{\sim} \text{PS}(\alpha)$$

where $\mu^*(\mathbf{s}) = \mu(\mathbf{s}) + \frac{\sigma(\mathbf{s})}{\xi(\mathbf{s})}\left\{\theta(\mathbf{s})^{\xi(\mathbf{s})} - 1\right\}$, $\sigma^*(\mathbf{s}) = \alpha\sigma(\mathbf{s})\theta(\mathbf{s})^{\xi(\mathbf{s})}$, and $\xi^*(\mathbf{s}) = \alpha\xi(\mathbf{s})$. MCMC sampling is performed with a Metropolis-within-Gibbs algorithm, and implementation of this model can be performed using the 'HKEVP' package in R.

A.3. Latent Variable Model

We use the latent variable model proposed by Davison et al. (2012) as an example to illustrate latent variable models for the GEV distribution. Davison et al. (2012) introduced a simple hierarchical structure, where the response variables $Y(\mathbf{s})$ are assumed to be independent conditionally on latent processes that describe the GEV parameters.

The hierarchical model is given by

$$\begin{aligned} Y(\mathbf{s})|\mu,\sigma,\xi &\overset{\text{iid}}{\sim} \text{GEV}\{\mu(\mathbf{s}),\sigma(\mathbf{s}),\xi(\mathbf{s})\},\\ \mu(\mathbf{s}) &= \boldsymbol{\beta}_\mu^T X(\mathbf{s}) + \epsilon_\mu(\mathbf{s}),\\ \sigma(\mathbf{s}) &= \boldsymbol{\beta}_\sigma^T X(\mathbf{s}) + \epsilon_\sigma(\mathbf{s}),\\ \xi(\mathbf{s}) &= \xi_0, \end{aligned}$$

where the mean function of $\mu(\mathbf{s})$ and $\sigma(\mathbf{s})$ are linear combinations of covariates $X(\mathbf{s})$ with coefficients $\boldsymbol{\beta}_\mu$ and $\boldsymbol{\beta}_\sigma$, respectively. Further, $\epsilon_\mu(\mathbf{s})$ and $\epsilon_\sigma(\mathbf{s})$ are stationary zero-mean Gaussian processes with correlation function $\rho_{\epsilon_\mu}(h) = \alpha_{\epsilon_\mu}\exp(-||h||/\lambda_{\epsilon_\mu})$ and $\rho_{\epsilon_\sigma}(h) = \alpha_{\epsilon_\sigma}\exp(-||h||/\lambda_{\epsilon_\sigma})$, respectively, and unknown sill and range parameters $\alpha_{\epsilon_\mu}, \alpha_{\epsilon_\sigma}, \lambda_{\epsilon_\mu}$ and $\lambda_{\epsilon_\sigma}$ (Sebille, Fougeres, and Mercadier 2016).

Using the same principals we adapt a simple latent Gaussian variable model as an example to illustrate latent variable models for the GPD distribution. The excesses above a threshold $u(\mathbf{s})$, $Y(\mathbf{s}) = X(\mathbf{s}) - u(\mathbf{s})|X(\mathbf{s}) > u(\mathbf{s})$, are assumed to be conditionally independent given

latent processes that describe the GPD parameters. The hierarchical model is given by

$$\begin{aligned} Y(\mathbf{s})|\tilde{\sigma},\xi &\overset{\text{indep}}{\sim} \text{GPD}\{\tilde{\sigma}(\mathbf{s}),\xi(\mathbf{s})\},\\ \log\{\tilde{\sigma}(\mathbf{s})\} &= \boldsymbol{\beta}_{\tilde{\sigma}}^T X(\mathbf{s}) + \epsilon_{\tilde{\sigma}}(\mathbf{s}),\\ \xi(\mathbf{s}) &= \xi_0, \end{aligned}$$

where the mean function of $\log\{\tilde{\sigma}(\mathbf{s})\}$ are linear combinations of covariates $X(\mathbf{s})$ with coefficients $\boldsymbol{\beta}_{\tilde{\sigma}}$ and $\epsilon_{\tilde{\sigma}}$ is a spatial Gaussian process, as for the GEV model.

Appendix B

A Taylor series is an expansion of a function into an infinite series of a variable or a finite series plus a remainder term. The coefficients of the expansion of the series involves taking successive derivatives of the function. The goal of Section 3.2 is to create an approximation of $l\{\mu(\mathbf{s}_i),\sigma(\mathbf{s}_i),\xi(\mathbf{s}_i);y(\mathbf{s}_i)\}$ that is quadratic with separable parameters. Putting a full second-order Taylor approximation with three parameters into quadratic form is extremely intricate, so we instead choose to estimate each parameter sequentially and focus on the likelihood approximation of each parameter separately. Consider the parameter $\mu(\mathbf{s})$, the second-order Taylor approximation about the current iterate $\{z_{\mu(\mathbf{s}_i)}, z_{\sigma(\mathbf{s}_i)}, z_{\xi(\mathbf{s}_i)}\}$ can be defined as:

$$\begin{aligned} l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\} \approx\ & l\{z_{\mu(\mathbf{s}_i)}, z_{\sigma(\mathbf{s}_i)}, z_{\xi(\mathbf{s}_i)}\}\\ &+\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\{\mu(\mathbf{s}_i)-z_{\mu(\mathbf{s}_i)}\}\\ &+\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\frac{\{\mu(\mathbf{s}_i)-z_{\mu(\mathbf{s}_i)}\}^2}{2}. \end{aligned}$$

We can ignore terms that are constant given $\{z_{\mu(\mathbf{s}_i)}, z_{\sigma(\mathbf{s}_i)}, z_{\xi(\mathbf{s}_i)}\}$ as they are irrelevant in minimization. Expand and simplify the relevant part of the likelihood:

$$\begin{aligned} &\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\{\mu(\mathbf{s}_i)-z_{\mu(\mathbf{s}_i)}\}\\ &\quad+\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\frac{\{\mu(\mathbf{s}_i)-z_{\mu(\mathbf{s}_i)}\}^2}{2}\\ &=\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\mu(\mathbf{s}_i)\\ &\quad-\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot z_{\mu(\mathbf{s}_i)}\\ &\quad+\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\frac{\mu(\mathbf{s}_i)^2}{2}\\ &\quad-\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\mu(\mathbf{s}_i)z_{\mu(\mathbf{s}_i)}\\ &\quad+\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\frac{z_{\mu(\mathbf{s}_i)}^2}{2}\\ &=\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot\frac{\mu(\mathbf{s}_i)^2}{2}\\ &\quad-\left\{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\cdot z_{\mu(\mathbf{s}_i)}\right.\\ &\quad\left.+\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}\right\}\mu(\mathbf{s}_i)+\text{constant}. \end{aligned}$$

Now the approximation can be written in quadratic form in terms of $\mu(\mathbf{s}_i)$ by completing the square

$$\begin{aligned} &=\frac{1}{2}\left[\sqrt{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}\right.\\ &\quad\times\left\{z_{\mu(\mathbf{s}_i)}-\frac{\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}\right\}\\ &\quad\left.-\sqrt{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}\cdot\mu(\mathbf{s}_i)\right]^2+\text{constant} \end{aligned}$$

Then in the least-square approximation, we can put

$$\begin{aligned} y &= \sqrt{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}\\ &\quad\times\left\{z_{\mu(\mathbf{s}_i)}-\frac{\frac{\partial l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}\right\}\quad\text{and}\\ x &= \sqrt{\frac{\partial^2 l\{\mu(\mathbf{s}_i);y(\mathbf{s}_i)\}}{\partial\mu(\mathbf{s}_i)^2}\Big|_{\mu(\mathbf{s}_i)=z_{\mu(\mathbf{s}_i)}}}. \end{aligned}$$

The same applies to $\sigma(\mathbf{s}_i)$ and $\xi(\mathbf{s}_i)$ of the GEV distribution and $\tilde{\sigma}(\mathbf{s}_i)$ and $\xi(\mathbf{s}_i)$ of the GPD distribution.

We can define the likelihood, first, and second derivatives of the GEV distribution as follows:

$$\begin{aligned} l(\mu,\sigma,\xi) &= -m\log(\sigma)-\left(1+\frac{1}{\xi}\right)\sum_{j=1}^{m}\log\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}\\ &\quad-\sum_{t=1}^{m}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}}\\ \frac{\partial l(\mu,\sigma,\xi)}{\partial\mu} &= \left(1+\frac{1}{\xi}\right)\sum_{t=1}^{m}\left\{\frac{\xi}{\sigma+\xi(y_t-\mu)}\right\}\\ &\quad-\sum_{t=1}^{m}\frac{1}{\sigma}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1}\\ \frac{\partial^2 l(\mu,\sigma,\xi)}{\partial\mu^2} &= \left(1+\frac{1}{\xi}\right)\sum_{t=1}^{m}\left[\frac{\xi^2}{\{\sigma+\xi(y_t-\mu)\}^2}\right]\\ &\quad-\frac{1+\xi}{\sigma^2}\sum_{t=1}^{m}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-2}\\ \frac{\partial l(\mu,\sigma,\xi)}{\partial\sigma} &= -\frac{m}{\sigma}+(1+\xi)\sum_{t=1}^{m}\left[\frac{y_t-\mu}{\sigma\{\sigma+\xi(y_t-\mu)\}}\right]\\ &\quad-\sum_{t=1}^{m}\frac{y_t-\mu}{\sigma^2}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1}\\ \frac{\partial^2 l(\mu,\sigma,\xi)}{\partial\sigma^2} &= \frac{m}{\sigma^2}-(1+\xi)\sum_{t=1}^{m}\left[\frac{y_t-\mu}{\sigma\{\sigma+\xi(y_t-\mu)\}^2}\right]\\ &\quad-(1+\xi)\sum_{t=1}^{m}\left[\frac{y_t-\mu}{\sigma^2\{\sigma+\xi(y_t-\mu)\}^2}\right]\\ &\quad-(1+\xi)\sum_{t=1}^{m}\left(\frac{y_t-\mu}{\sigma^2}\right)^2\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-2}\\ &\quad+2\sum_{t=1}^{m}\frac{y_t-\mu}{\sigma^3}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1}, \end{aligned}$$

$$\begin{aligned}\frac{\partial l(\mu,\sigma,\xi)}{\partial \xi} &= -\left(1+\frac{1}{\xi}\right)\sum_{t=1}^{m}\left[\frac{y_t-\mu}{\{\sigma+\xi(y_t-\mu)\}}\right]\\ &+\frac{1}{\xi}\sum_{t=1}^{m}\left(\frac{y_t-\mu}{\sigma}\right)\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1}\\ &+\frac{1}{\xi^2}\sum_{t=1}^{m}\log\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}\\ &\times\left[1-\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}}\right],\\ \frac{\partial^2 l(\mu,\sigma,\xi)}{\partial \xi^2} &= \sum_{t=1}^{m}\frac{2(y_t-\mu)}{\xi^2\{\sigma+\xi(y_t-\mu)\}}+\left(1+\frac{1}{\xi}\right)\\ &\times\sum_{t=1}^{m}\frac{(y_t-\mu)^2}{\{\sigma+\xi(y_t-\mu)\}^2}\\ &-\frac{2}{\xi^3}\sum_{t=1}^{m}\log\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}\\ &-\sum_{t=1}^{m}\frac{y_t-\mu}{\xi^2\{\sigma+\xi(y_t-\mu)\}}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-1/\xi}\\ &+\frac{2}{\xi^3}\sum_{t=1}^{m}\log\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}}\\ &-\frac{1}{\xi^4}\sum_{t=1}^{m}\left[\log\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}\right]^2\\ &\times\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}}\\ &+\sum_{t=1}^{m}\frac{2(y_t-\mu)}{\sigma\xi^3}\log\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}\\ &\times\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1}\\ &-\left(1+\frac{1}{\xi}\right)\sum_{t=1}^{m}\frac{(y_t-\mu)^2}{\sigma^2\xi}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-2}\\ &-\sum_{t=1}^{m}\frac{y_t-\mu}{\sigma\xi^2}\left\{1+\xi\left(\frac{y_t-\mu}{\sigma}\right)\right\}^{-\frac{1}{\xi}-1}.\end{aligned}$$

We can define the likelihood, first, and second derivatives of the GPD distribution as follows:

$$\begin{aligned}l(\tilde{\sigma},\xi) &= -k\log\tilde{\sigma}-\left(1+\frac{1}{\xi}\right)\sum_{t=1}^{k}\log\left(1+\xi\frac{y_t}{\tilde{\sigma}}\right),\\ \frac{\partial l(\tilde{\sigma},\xi)}{\partial \tilde{\sigma}} &= \frac{k}{\tilde{\sigma}\xi}-\left(1+\frac{1}{\xi}\right)\sum_{t=1}^{k}\frac{1}{\tilde{\sigma}+\xi y_t},\\ \frac{\partial^2 l(\tilde{\sigma},\xi)}{\partial \tilde{\sigma}^2} &= -\frac{k}{\tilde{\sigma}^2\xi}+\left(1+\frac{1}{\xi}\right)\sum_{t=1}^{k}\frac{1}{(\tilde{\sigma}+\xi y_t)^2},\\ \frac{\partial l(\tilde{\sigma},\xi)}{\partial \xi} &= -\frac{k}{\xi^2}(1+\log\tilde{\sigma})-\frac{k}{\xi}\\ &+\frac{1}{\xi^2}\sum_{t=1}^{k}\log(\tilde{\sigma}+\xi y_t)+\left(1+\frac{1}{\xi}\right)\frac{\tilde{\sigma}}{\xi}\sum_{t=1}^{k}\frac{1}{\tilde{\sigma}+\xi y_t},\\ \frac{\partial^2 l(\tilde{\sigma},\xi)}{\partial \xi^2} &= \frac{2k}{\xi^3}(1.5+\log\tilde{\sigma})+\frac{k}{\xi^2}\\ &-\frac{2}{\xi^3}\sum_{t=1}^{k}\log(\tilde{\sigma}+\xi y_t)-\frac{\tilde{\sigma}(\xi+3)}{\xi^3}\sum_{t=1}^{k}\frac{1}{\tilde{\sigma}+\xi y_t}\\ &-\frac{\tilde{\sigma}}{\xi}\left(1+\frac{1}{\xi}\right)\sum_{t=1}^{k}\frac{y_t}{(\tilde{\sigma}+\xi y_t)^2}.\end{aligned}$$

Appendix C

Table C1 and C2 show the model generating configurations for the stationary and nonstationary data, respectively. The max-stable field is generated using the Schlather model. The GEV parameters and covariance function in Table C2 are scaled by a factor of 20 to adjust for the scaled spatial domain $\mathcal{D} = [0,1]^2$, providing equivalent parameters to those in Table C1.

Appendix D

The following tables provide supplementary information for the simulation studies conducted in Section 4. This includes the average RMSE for the parameter and return level estimates, as well as the average computation time for the 200 simulations in each simulation study.

Table C1. Generating model configurations used to simulate stationary data for the model comparisons in Sections 4.2.1 and 4.3.1.

Spatial sample size	$N \in 200$ sites sampled uniformly on $\mathcal{D} = [0,20]^2$
Temporal sample size	$T \in (50, 500)$ for GEV and GPD respectively
Max-stable field settings	Covariance = Whittle-Matern Range = 1 Smoothness = .5 Nugget = 0
Distributions for GEV parameters $\theta(\mathbf{s})$	Covariance function $\rho_{(\sigma_0,\lambda_0,\nu_0)}(\mathbf{s}_1,\mathbf{s}_2) = \sigma_0\exp\{-(\|\mathbf{s}_1-\mathbf{s}_2\|/\lambda_0)^{\nu_0}\}$ Gaussian processes $\mu(\mathbf{s}) \sim GP\{26 + [.5 \quad 0]^T\mathbf{s}, \rho_{(4,20,1)}\}$ $log\{\sigma(\mathbf{s})\} \sim GP\{log(10) + [0 \quad .05]^T\mathbf{s}, \rho_{(.4,5,1)}\}$ $\xi(\mathbf{s}) \sim GP\{.12, \rho_{(.0012,10,1)}\}$

Table C2. Generating model configurations used to simulate nonstationary data for the model comparisons in Sections 4.2.2 and 4.3.2.

Spatial sample size	$N \in 200$ sites sampled uniformly on $\mathcal{D} = [0,1]^2$
Temporal sample size	$T \in (50, 500)$ for GEV and GPD respectively
Max-stable field settings	Covariance = Whittle-Matern Range = 1 Smoothness = .5 Nugget = 0
Distributions for GEV parameters $\theta(\mathbf{s})$	Covariance function $\rho_{(\sigma_0,\lambda_0,\nu_0)}(\mathbf{s}_1,\mathbf{s}_2) = \sigma_0\exp\{-(\|\mathbf{s}_1-\mathbf{s}_2\|/\lambda_0)^{\nu_0}\}$ Gaussian processes $\mu(\mathbf{s}) \sim GP\{26 + [10 \quad 0]^T\mathbf{s}, \rho_{(4,1,1)}\}$ $log\{\sigma(\mathbf{s})\} \sim GP\{log(10) + [0 \quad 1]^T\mathbf{s}, \rho_{(.4,.25,1)}\}$ $\xi(\mathbf{s}) \sim GP\{.12, \rho_{(.0012,.5,1)}\}$

NOTE: The GEV parameters and covariance function are scaled by a factor of 20 compared to Table C1 to provide equivalent parameters with the scaled spatial domain $\mathcal{D} = [0,1]^2$.

Table D1. The average RMSE for GEV parameter and return level estimates in the stationary simulation study in Section 4.2.1.

	Spatial GEV	Schlather	HKEVP	LVM	Fused ridge	Fused lasso
Location	2.111	2.079	1.364	1.432	2.475	3.606
Scale	14.298	13.474	2.951	3.246	8.903	10.716
Shape	0.1050	0.0955	0.1068	0.0581	0.0693	0.0905
20-year RL	48.146	45.578	19.270	14.926	35.609	42.418
50-year RL	66.874	63.217	34.704	23.156	51.512	60.867
100-year RL	82.787	78.194	51.539	31.102	65.670	77.253
Comp. time	1.79	61.11	15, 693.71	345.15	4.32	1.90

Note: The computation time is displayed in seconds.

Table D2. The average RMSE for GEV parameter and return level estimates in the nonstationary simulation study in Section 4.2.2.

	Spatial GEV	Schlather	HKEVP	LVM	Fused ridge	Fused lasso
Location	2.669	2.442	1.858	2.126	3.150	3.973
Scale	13.151	13.730	5.471	3.558	8.781	10.203
Shape	0.1136	0.0960	0.1476	0.0845	0.0915	0.1132
20-year RL	45.651	46.469	36.309	17.764	36.779	42.200
50-year RL	64.126	64.520	65.147	28.116	53.887	61.659
100-year RL	80.063	79.844	98.178	38.246	69.370	79.503
Comp. time	1.96	77.87	15, 549.65	346.74	4.77	2.15

Note: The computation time is displayed in seconds.

Table D3. The average RMSE for GPD parameter and return level estimates in the stationary simulation study in Section 4.3.1.

	Spatial GPD	LVM	Fused ridge	Fused lasso
Scale	14.553	5.699	8.394	13.167
Shape	0.1320	0.1660	0.0915	0.1834
20-year RL	14.321	4.517	6.618	9.547
50-year RL	27.795	8.505	14.966	21.841
100-year RL	40.369	13.718	22.184	32.091
Comp. time	1.32	635.36	13.80	0.86

Note: The computation time is displayed in seconds.

Table D4. The average RMSE for GPD parameter and return level estimates in the nonstationary simulation study in Section 4.3.2.

	Spatial GPD	LVM	Fused ridge	Fused lasso
Scale	14.902	6.110	8.835	13.158
Shape	0.1331	0.1649	0.0953	0.1805
20-year RL	14.723	4.721	6.871	9.590
50-year RL	28.941	8.878	15.894	22.243
100-year RL	42.528	14.181	23.897	33.158
Comp. time	2.09	620.83	14.49	1.20

Note: The computation time is displayed in seconds.

Appendix E

We conduct a simulation study comparing independent parameter estimates and sequential parameter estimates. Figure E1 shows that for the fused ridge and fused lasso model, the average RMSE is very similar when parameters are estimated sequentially or independently. The sequential estimation appears to be best when $\sigma(\mathbf{s})$ is estimated before $\xi(\mathbf{s})$. Based on the results, any estimation technique for the fused ridge model results in an improvement over the Spatial GEV. The fused lasso model produces very similar results to the Spatial GEV. Since the parameters we generated for the data is not in the form of clusters, it is not surprising that the fused ridge performs better than the fused lasso.

Because the accuracy of the Taylor approximation is dependent on the values chosen to expand about, the closer the value is to the true value the more accurate the results, it is expected that the average RMSE for the parameters would increase when estimated independently compared to being estimated sequentially. Following this notion, for the simulation study in Section 4, we choose to estimate parameters in the order of $\sigma(\mathbf{s}) \rightarrow \xi(\mathbf{s}) \rightarrow \mu(\mathbf{s})$, because $\sigma(\mathbf{s})$ shows the greatest improvement over the Spatial GEV allowing for better initialization when estimating $\xi(\mathbf{s})$, and so forth. However, the results also show that for the fused ridge model estimating parameters independently is still a significant improvement over the spatial GEV model, and only a relatively small change compared to the sequential estimation making it a reasonable alternative in applications where sequential estimation is not practical.

Appendix F

The following figures provide supplementary information to the two data applications conducted in Section 5. Figure F1 contains the corresponding standard errors to the estimated GEV parameters, 20-, and 100-year return level for the historical simulations in Figure 5 of Section 5.1. The change in estimated parameters and return levels between the historical simulation and future conditions simulation are in Figure F2 and the standard errors for those estimates are in Figure F3. While the standard errors in the 100-year return level may appear large, our block bootstrap simulation study in Appendix G shows that the bootstrap procedure actually tends to underestimate the actual uncertainty if the true coefficient surfaces are nonsmooth and captures the actual uncertainty if the true coefficient surfaces are smooth.

Finally, Figure F4 provides the nonparametric block bootstrap standard errors for the temperature estimates in Figure 6 of Section 5.2. The corresponding parameter estimates with their standard errors and the changes in parameter estimates with their standard errors are in F5 and F6.

Appendix G

We use block bootstrap to construct a 95% confidence interval for the parameters, 20-, 50-, and 100-year return level with our fused ridge and fused lasso model using both stationary and nonstationary data and compare the results to the LVM coverage. We also compare the empirical coverage probability (ECP) of the 95-percent confidence intervals to its nominal level. If the ECP is close to 95-percent, then the uncertainty of the estimation is appropriately quantified.

Figure G1 shows the ECP for the GEV parameters and 20-, 50-, and 100-year return levels of 200 simulations with 200 bootstrap samples. The data is generated using the simulation settings in Appendix C for stationary and nonstationary GEV data. Figure G1 shows the fused models have better coverage for the location and shape parameter while the LVM model has better coverage for the scale parameter. The LVM model for stationary data performs better in terms of return level coverage compared to the fused models, however the fused models are closer to the nominal level with nonstationary data.

The fused models have undercoverage for the scale parameter because the generated true scale parameter surface is not as smooth as the other two parameters, and the fused penalties impose smooth estimates. To demonstrate the coverage of the fused models when the true parameter surfaces are smooth functions of space, we modify the data generation in Appendix C such that the covariance function of $log\{\sigma(\mathbf{s})\}$ is $\rho_{(.4,15,1)}$ for the stationary data and $\rho_{(.4,.75,1)}$ for the nonstationary data (compared to $\rho_{(.4,5,1)}$ and $\rho_{(.4,.25,1)}$, respectively). Figure G2 shows the resulting ECP for the LVM and fused models under a smoothly generated scale parameter. The resulting scale parameter coverage is now significantly closer to the nominal level and the return level coverage is at the nominal level for both stationary and nonstationary data using our fused models. The LVM model shows undercoverage for both the parameters and return levels with results very similar to Figure G1.

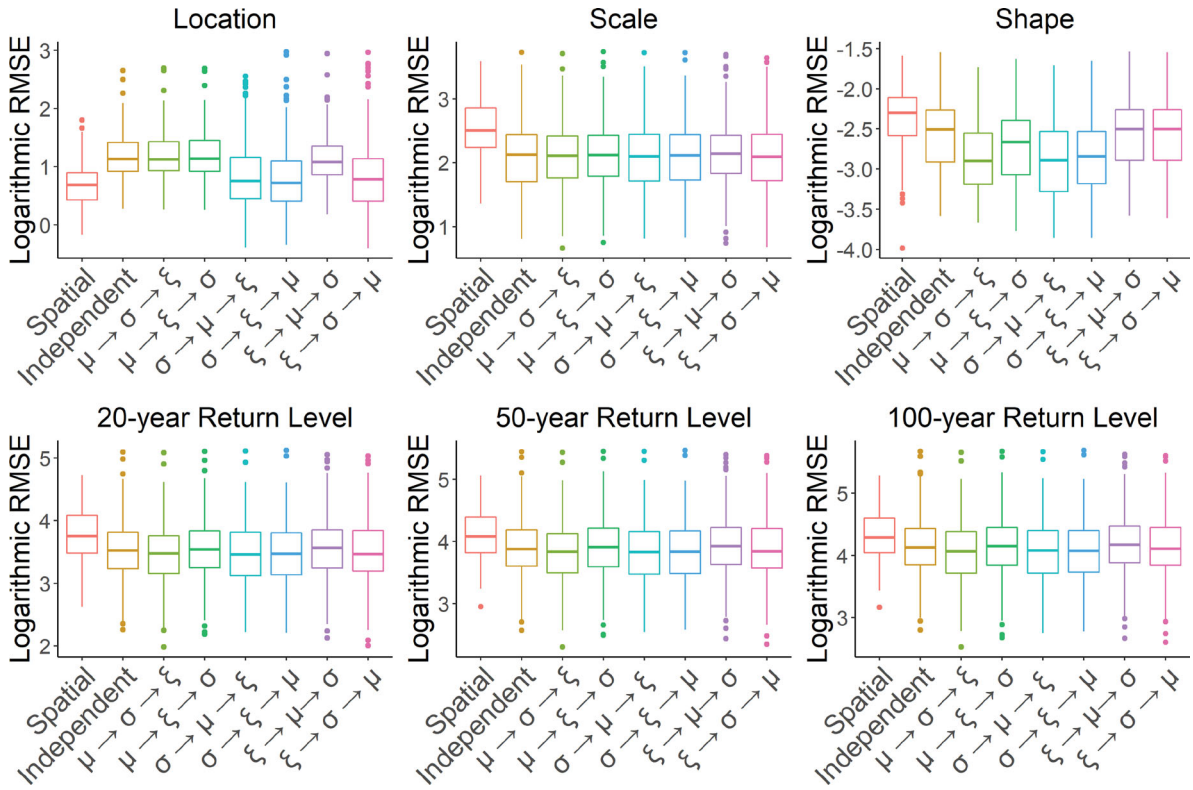

Figure E1. Boxplots of logarithmic RMSE for the GEV parameters and return level estimates with stationary spatial extremes and eight models.

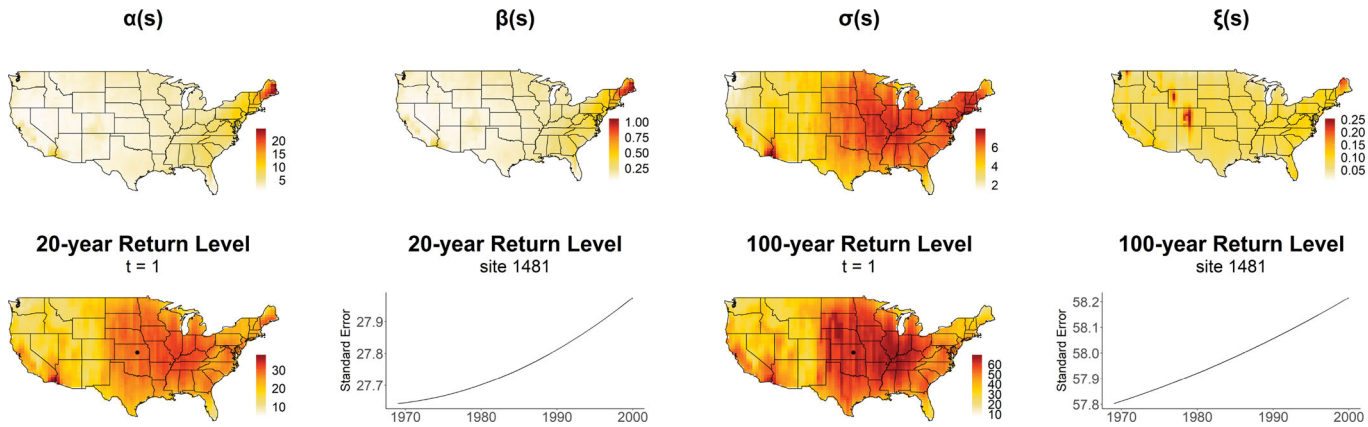

Figure F1. Standard error of the GEV parameters, 20-, and 100-year return level for the historical simulation using the fused ridge model. Since the return level is time varying, the spatially varying return level is shown at time t=1 (year 1969) and the standard error for site 1481, located at the black dot, demonstrates how the return level varies over time. All units are mm/h.

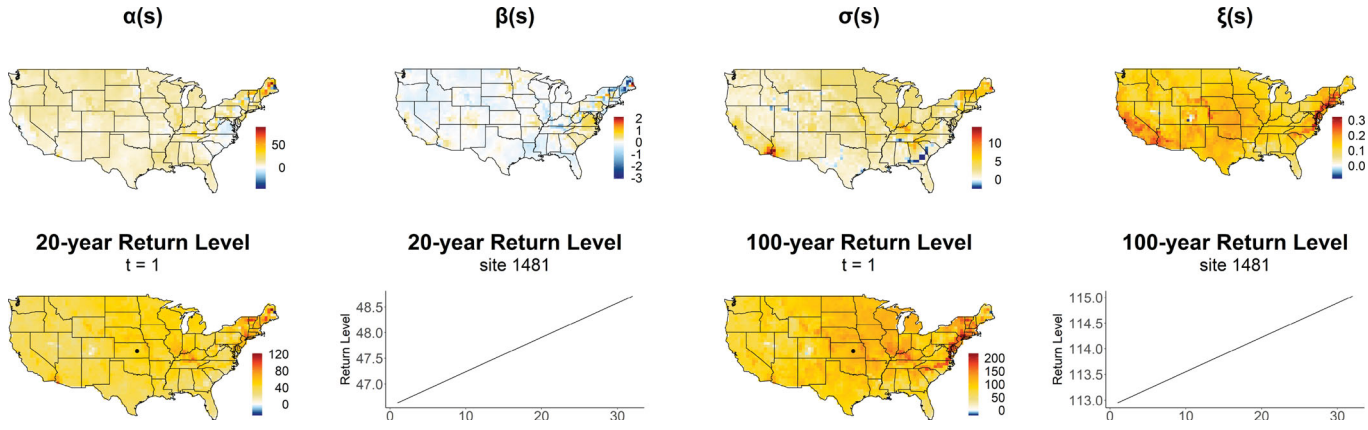

Figure F2. Estimated change from historical to future conditions simulation for the fused ridge model. All units are mm/h.

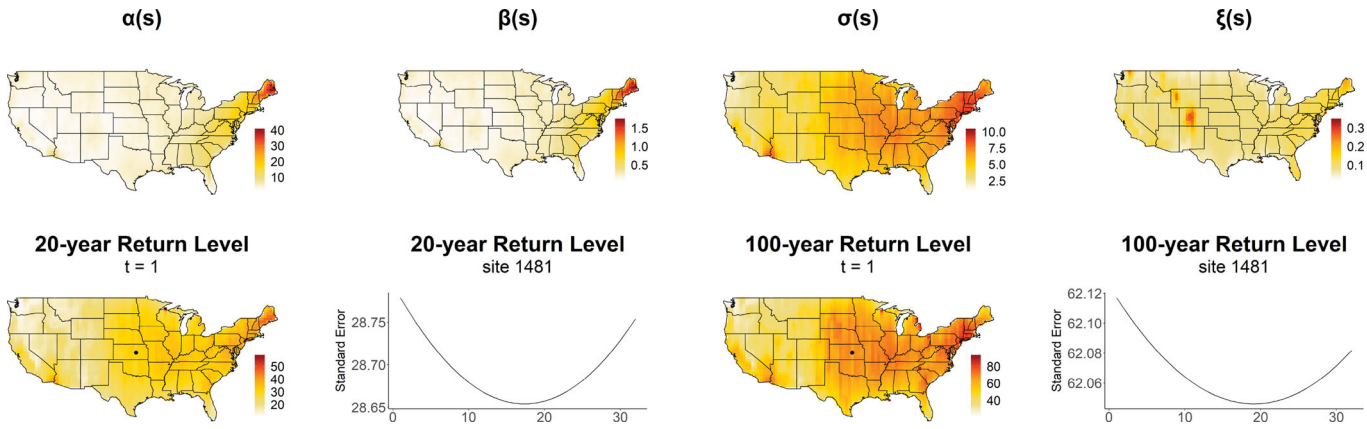

Figure F3. Estimated standard error of the change from historical to future conditions simulation for the fused ridge model. All units are mm/h.

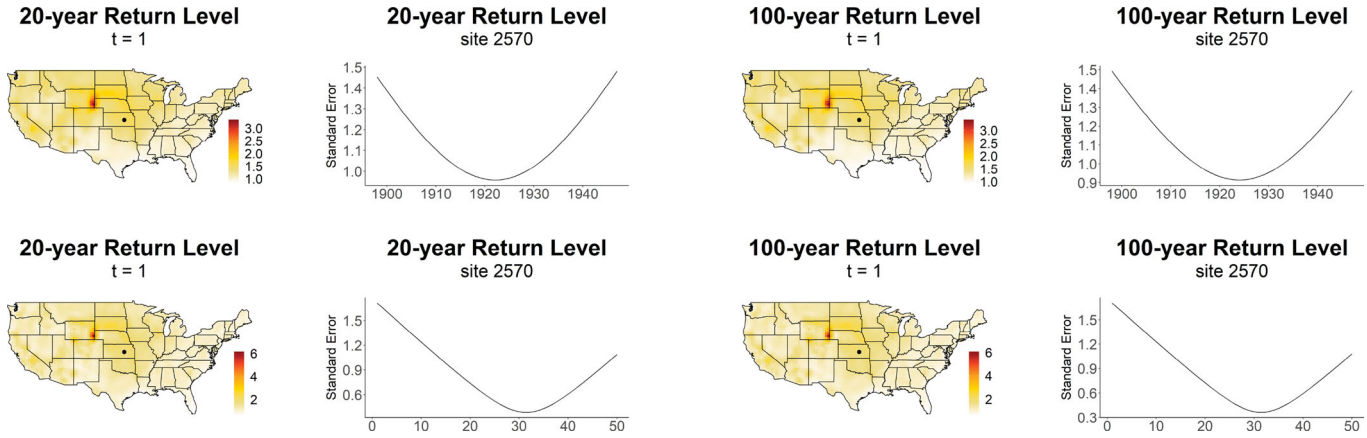

Figure F4. Top: Standard error of the 20- and 100-year return level for the years 1898–1947 using the fused ridge model. Bottom: Standard error of the change in 20- and 100-year return level between the years 1898–1947 and 1948–1997 using the fused ridge model. The spatially varying return levels are shown at time t=1 (1898 and 1948) and the time varying return levels are shown for site 2570, located at the black dot. All units are in degrees Celsius.

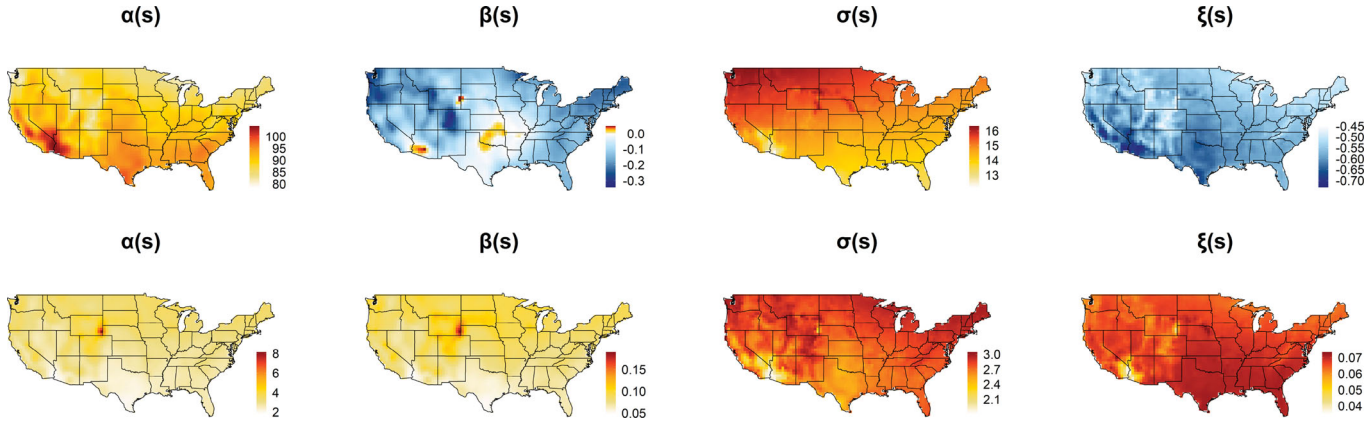

Figure F5. Top: Estimated GEV parameters for the years 1898–1947 using the fused ridge model. Bottom: Standard error of the GEV parameter for the years 1898-1947 using the fused ridge model. All units are in degrees Fahrenheit. Return levels were converted to degrees Celsius after estimation.

The simulation study shows that for our fused models this bootstrap procedure has empirical coverage near the nominal level when the true parameter surfaces are smooth functions of space, but undercoverage when the true parameter surfaces are not sufficiently smooth. The simulation also shows that in both cases the LVM has undercoverage.

Figures G3 and G4 demonstrates the GPD parameters, 20-, 50-, and 100-year return level coverage of 200 simulations with 200 bootstrap samples for stationary and nonstationary GPD spatial extremes. Figure G3 uses the data-generation settings from Appendix C while Figure G4 uses the modified scale parameter. We cannot use the smoothed threshold for the bootstrap study due to an incompatibility with the "smoothr" package and our computing cluster, however we hope to show that the smoothed $\tilde{\sigma}$ parameter also helps to ease the undercoverage of CIs with a GPD model. The overcoverage of the fused ridge in Figure G3 and G4 may be alleviated when the smoothed threshold is applied. A small simulation with a smoothed threshold showed that ECP of return levels with fused ridge and fused lasso are both slightly below the nominal level when the $\tilde{\sigma}$ parameter is not sufficiently smooth. The coverage of the shape parameter in the GPD model seems unsatisfactory compared to that of the return levels. Even

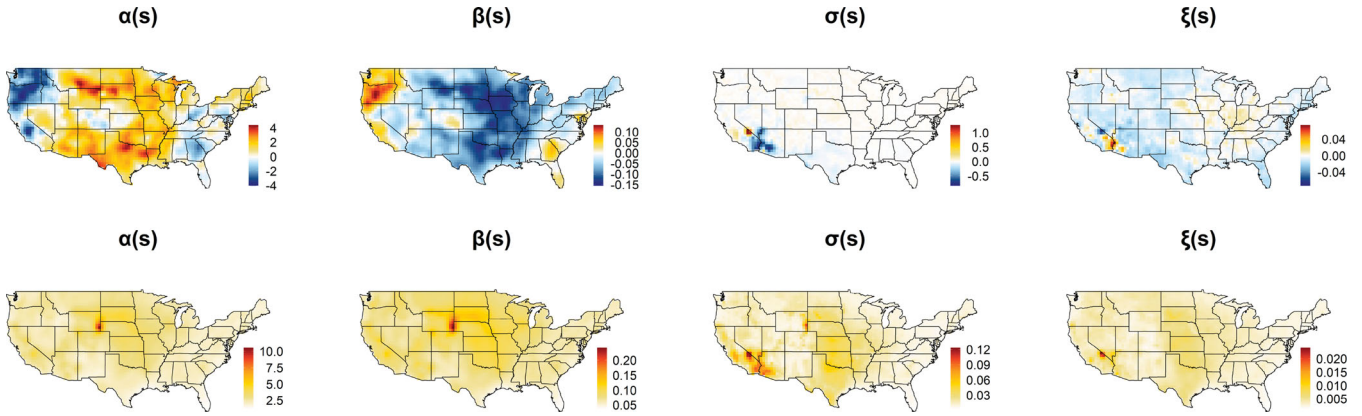

Figure F6. Top: Estimated change in GEV parameters between the years 1898–1947 and 1948–1997 using the fused ridge model. Bottom: Standard error of the change in GEV parameter estimates between the years 1898–1947 and 1948–1997 using the fused ridge model. All units are in degrees Fahrenheit. Return levels were converted to degrees Celsius after estimation.

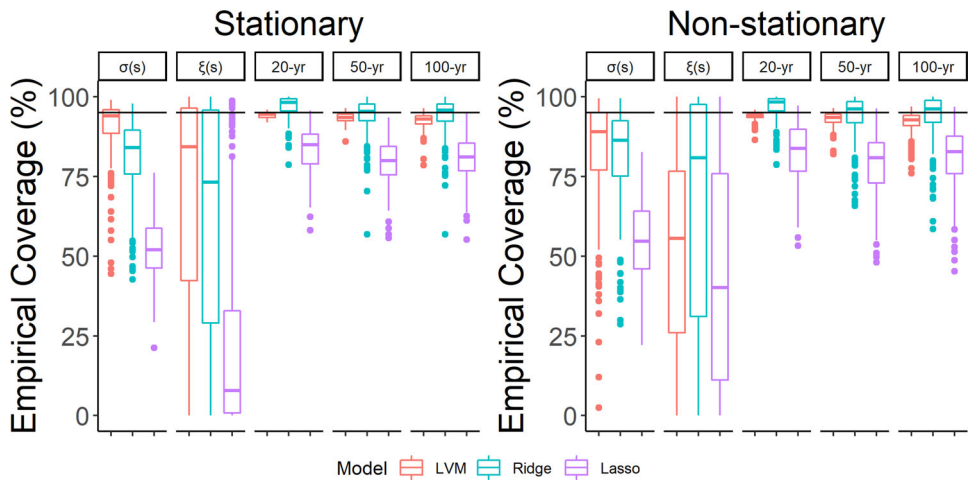

Figure G1. Empirical coverage probability (ECP) of the 95% confidence interval for stationary and nonstationary spatial extremes using three GEV models: Latent Variable, Fused spatial GEV under Ridge and Lasso penalty. The data are generated using the settings in Appendix C. The solid line represents the nominal level of 95%.

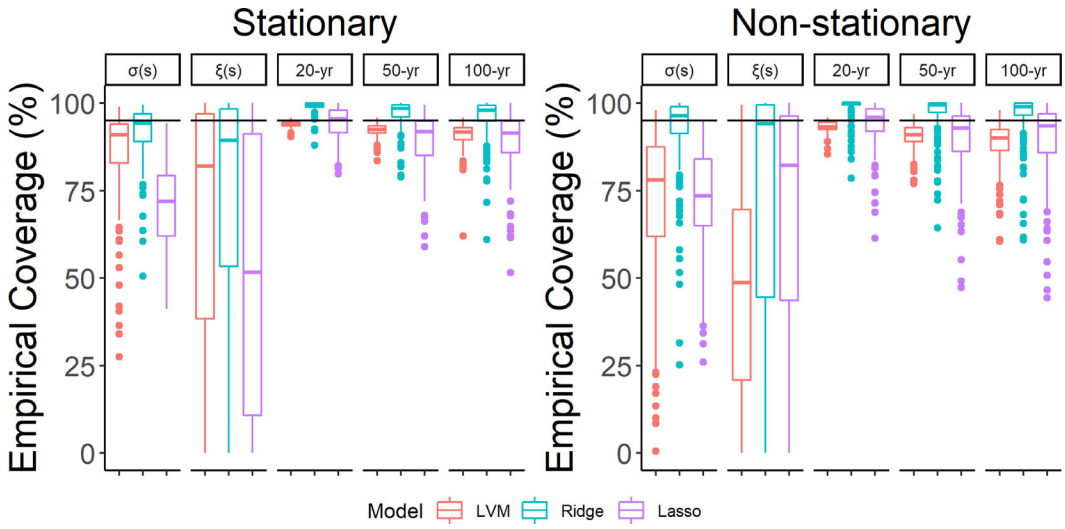

Figure G2. ECP of the 95-percent confidence interval for stationary and nonstationary spatial extremes using three GEV models: Latent Variable, Fused spatial GEV under Ridge and Lasso penalty. The data generation is modified such that the true parameters are smooth functions of space. The solid line represents the nominal level of 95%.

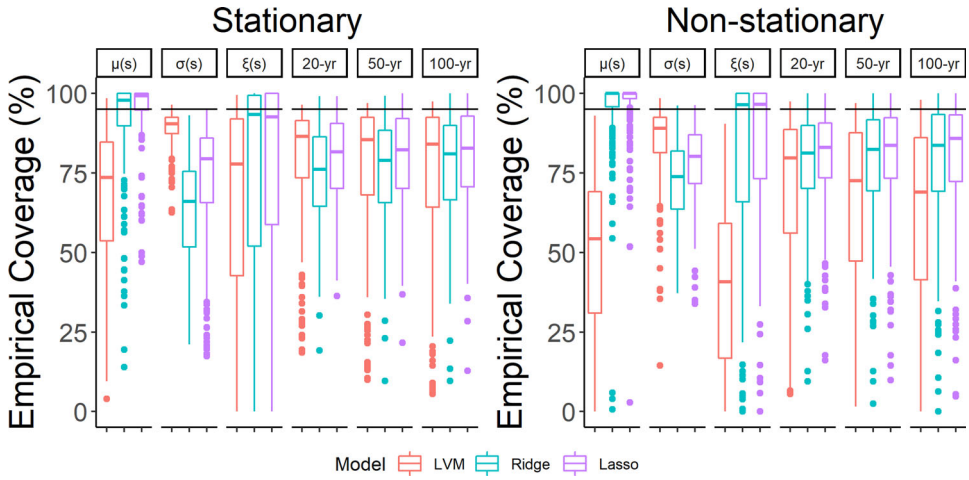

Figure G3. ECP of the 95-percent confidence interval for stationary and nonstationary spatial extremes using three GPD models: Latent Variable, Fused spatial GEV under Ridge and Lasso penalty. The data are generated using the settings in Appendix C. The solid line represents the nominal level of 95%.

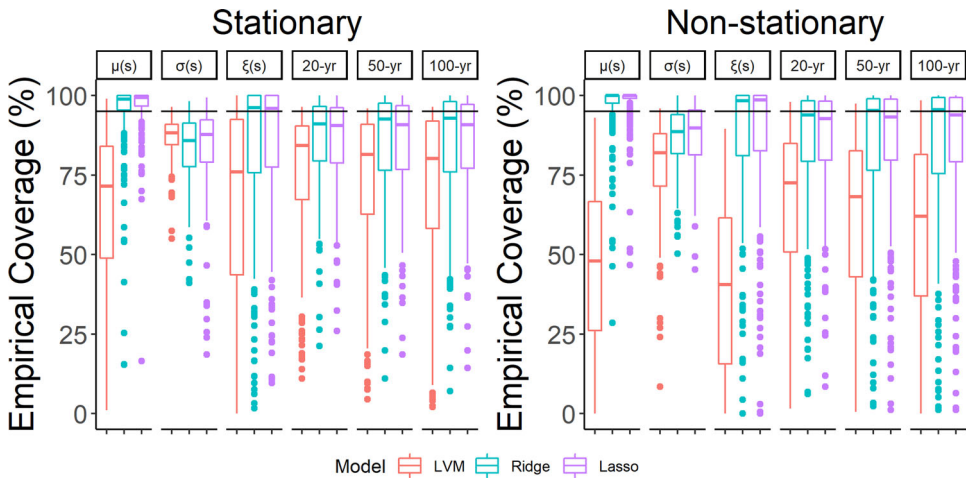

Figure G4. ECP of the 95% confidence interval for stationary and nonstationary spatial extremes using three GPD models: Latent Variable, Fused spatial GEV under Ridge and Lasso penalty. The data generation is modified such that the true parameters are smooth functions of space. The solid line represents the nominal level of 95%.

the LVM shows no advantage over the fused ridge models in improving the coverage of the shape parameter. Further investigation is needed to determine the reason in order to propose a remedial approach for this issue.

References

Bopp, G., Shaby, B., and Huser, R. (2021), "A Hierarchical Max-Infinitely Divisible Spatial Model for Extreme Precipitation," *Journal of the American Statistical Association*, 116, 93–106. [1]

Burrough, P., and McDonnell, R. (1998), *Principles of Geographical Information Systems*, Oxford: Oxford University Press. [4]

Cao, Y., and Li, B. (2018), "Assessing Models for Spatial Extremes and Methods for Uncertainty Quantification on Return Level Estimation," *Environmetrics*. [3,5,9]

Casson, E., and Coles, S. (1999), "Spatial Regression Models for Extremes," *Extremes*, 1, 449–468. [1]

Castro Camilo, D.. and Huser, R. (2020), "Local Likelihood Estimation of Complex Tail Dependence Structures, Applied to U.S. Precipitation Extremes," *Journal of the American Statistical Association*, 115, 1037–1054. [1]

Cheng, R., and Amin, N. (1983), "Estimating Parameters in Continuous Univariate Distributions With a Shifted Origin," *Journal of the Royal Statistical Society*, Series B, 45, 394–403. [4]

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, London, UK: Springer-Verlag London. [1,2,5,7]

Coles, S., and Casson, E. (1998), "Extreme Value Modelling of Hurricane Wind Speeds," *Structural Safety*, 20, 283–296. [1]

Coles, S., and Tawn, J. (1996), "A Bayesian Analysis of Extreme Rainfall Data," *Journal of the Royal Statistical Society*, Series C, 45, 463–478. [2]

Cooley, D., Nychka, D., and Naveau, P. (2007), "Bayesian Spatial Modeling of Extreme Precipitation Return Levels," *Journal of the American Statistical Association*, 102, 824–840. [1,7]

Davis, R., and Resnik, S. (1993), "Prediction of Stationary Max-Stable Processes," *The Annals of Applied Probability*, 3, 497–525. [1]

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, New York: Cambridge University Press. [5]

Davison, A. C., Huser, R., and Thibaud, E. (2019), Spatial Extremes, in *Handbook of Environmental and Ecological Statistics*, eds A. E. Gelfand, M. Fuentes, J. A. Hoeting and R. L. Smith, Boca Raton, FL: CRC Press. [2]

Davison, A. C., Padoan, S. A., and Ribatet (2012), "Statistical Modeling of Spatial Extremes," *Statistical Science*, 27, 161–186. [2,6,11]

Davison, A. C., and Smith, R. L. (1990), "Models for Exceedances Over High Thresholds," *Journal of the Royal Statistical Society*, Series B, 52, 393–442. [1]

de Haan, L. (1984), "A Spectral Representation for Max-Stable Processes," *The Annals of Probability*, 12, 1194–1204. [11]

Fisher, R. A., and Tippett, L. H. C. (1928), "On the Estimation of the Frequency Distributions of the Largest or Smallest Member of a Sample,"

Proceedings of the Cambridge Philosophical Society, 24, 180–190. [1]

Golub, G. H., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223. [4]

Gumbel, E. J. (1958), *Statistics of Extremes*, New York: Columbia University Press. [1]

Hewitt, J., Fix, M., Hoeting, J., and Cooley, D. (2019), "Improved Return Level Estimation Via a Weighted Likelihood, Latent Spatial Extremes Model," *Journal of Agricultural Biological and Environmental Statistics*, 24, 426–443. [5]

Huser, R., Opitz, T., and Thibaud, E. (2017), "Bridging Asymptotic Independence and Dependence in Spatial Extremes Using Gaussian Scale Mixtures," *Spatial Statistics*, 21, 166–186. [1]

——— (2021), "Max-Infinitely Divisible Models and Inference for Spatial Extremes," *Scandinavian Journal of Statistics 48*(1), 321–348. [1]

Huser, R., and Wadsworth, J. L. (2019), "Modeling Spatial Processes with Unknown Extremal Dependence Class," *Journal of the American Statistical Association*, 114, 434–444. [1]

Johns, C., Nychka, D., Kittel, T., and Daly, C. (2003), "Infilling Sparse Records of Spatial Fields," *Journal of the American Statistical Association*, 98, 796–806. [10]

Lee, J., Li, S., and Lund, R. (2014), "Trends in Extreme U.S. Temperatures," *Journal of Climate*, 27, 4209–4225. [10]

Li, F., and Sang, H. (2019), "Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets," *Journal of the American Statistical Association*, 114, 1050–1062. [3,4]

Lindsay, B. G. (1988), "Composite Likelihood Methods," *Contemporary Mathematics*, 80, 221–239. [2]

Longley, P., Goodchild, M., Maguire, D., and Rhind, D. (2001), *Geographic Information Systems and Science*. Chichester: Wiley. [4]

Lund, R., Seymour, L., and Kafadar, K. (2001), "Temperature Trends in the United States," *Environmetrics,* 12, 673–690. [10]

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Its Applications*, 9, 141–142. [7]

——— (1989), *Nonparametric Estimation of Probability Densities and Regression Curves*, Netherlands: Springer. [7]

Nakicenovic, N., Alcamo, J., Davis, G., and De Vries, B. (2000), *Special Report on Emissions Scenarios: A Special Report of Working Group III of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press. [8]

Padoan, S. A., Ribatet, M., and Sisson, S. A. (2010), "Likelihood-Based Inference for Max-Stable Processes," *Journal of the American Statistical Association*, 105, 263–277. [2]

Parker, R. J., Reich, B. J., and Eidsvik, J. (2016), "A Fused Lasso Approach to Nonstationary Spatial Covariance Estimation," *Journal of Agricultural, Biological, and Environmental Statistics*, 21, 569–587. [3]

Pickands, J. (1975), "Statistical Inference Using Extreme Order Statistics," *Annals of Statistics*, 3, 119–131. [1]

Reich, B., and Shaby, B. (2012), "A Hierarchical Max-Stable Spatial Model for Extreme Precipitation," *The Annals of Applied Statistics*, 6, 1430–1451. [2,9,11]

Sang, H., and Gelfand, A. E. (2009), "Hierarchical Modeling for Extreme Values Observed Over Space and Time," *Environmental and Ecological Statistics*, 16, 407–426. [1]

Sang, H., and Genton, M. G. (2014), "Tapered Composite Likelihood for Spatial Max-Stable Models," *Spatial Statistics*, 8, 86–103. [2]

Schlather, M. (2002), "Models for Stationary Max-Stable Random Fields," *Extremes*, 5, 33–44. [1,2,11]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [4]

Sebille, Q., Fougeres, A., and Mercadier (2016), "A Comparison of Spatial Extreme Value Models. Application to Precipitation Data," Unpublished manuscript. [11]

Smith, R. L. (1985), "Maximum Likelihood Estimation in a Class of Non-regular Cases," *Biometrika*, 72, 67–92. [4]

——— (1990), "Max-Stable Processes and Spatial Extremes," Unpublished manuscript. [1]

Tansey, W., Koyejo, R. P., and Scott, J. (2018), "False Discovery Rate Smoothing," *Journal of the American Statistical Association*, 113, 1156–1171. [3,4]

Tibshirani, R. J., and Taylor, J. (2011), "The Solution Path of the Generalized Lasso," *Annals of Statistics*, 39, 1335–1371. [3]

Tye, M. R., and Cooley, D. (2015), "A Spatial Model to Examine Rainfall Extremes in Colorado's Front Range," *Journal of Hydrology*, 530, 15–23. [5]

Varin, C., and Vidoni, P. (2005), "A Note on Composite Likelihood Inference and Model Selection," *Biometrika*, 92, 519–528. [2,4]

Wadsworth, J. L., and Tawn, J. A. (2012), "Dependence Modelling for Spatial Extremes," *Biometrika*, 99, 253–272. [1]

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhya: The Indian Journal of Statistics*, Series A, 26, 359–372. [7]